JOURNAL FOR DERADICALIZATION

Summer 2025
No. 43
ISSN: 2363-9849

Examining the Impacts of a Gamified Media Literacy Intervention in Indonesia

Michael H. Becker[a1], Michael J. Williams[b], Alexa Hassan[c]
[a]Research Associate, University of Nebraska - Omaha, [b]Director, The Science of P/CVE, [c]Manager, Moonshot

Abstract
Integrating popular messaging applications and gamified approaches is an emerging strategy to deploy media literacy interventions at-scale. The present study examines the impacts of a WhatsApp-styled intervention in prebunking mis- and dis-information among an age and gender stratified sample of 504 Indonesian adults. Guided by the Theory of Planned Behavior, the intervention aimed to inoculate participants to common disinformation tactics and arguments through interactive elements simulating group chats with loved ones around three major topic areas (health, news and finance). Data were collected on participants' attitudes, subjective norms, self-efficacy, intentions, and ability to detect disinformation before and after the intervention. Following the intervention, participants reported a statistically significant increase in motivation to detect fake information, perception that it is normal to identify disinformation online, and greater motivation to combat misinformation. These impacts were sustained two weeks after the intervention. While self-efficacy increased immediately after the intervention, it returned to baseline levels at the two-week follow-up, despite this initial increase. Findings provide valuable insights into the potential of gamified interventions to effectively influence key behavioral determinants related to disinformation consumption and sharing, particularly motivation and perceived norms.

Article History
Received Feb 20, 2025
Accepted Jun 25, 2025
Published Jun 27, 2025

Keywords: Disinformation, Misinformation, Inoculation, Theory of Planned Behavior, Gamification

Introduction

In a 2024 address to the World Economic Forum, then President of the European Commission warned that "One of the greatest challenges we face in the digital age is the spread of misinformation. It erodes trust, fuels division, and undermines our ability to make informed decisions." (World Economic Forum, 2024). Evidence abounds of the human impacts of misinformation in recent years including reduced uptake of preventative medical treatments (Wang et al., 2019), welfare losses (Bairoliya & McKiernan, 2024), reduced security in the

[1] Corresponding Author Contact: Michael H. Becker, Email: michael.h.becker1@gmail.com; ORCID: https://orcid.org/0000-0002-0168-1499

wake of natural disasters (Rascoe, 2024), and more. Indeed, mis- and dis-information have been observed to reinforce ethnic and political identities, rendering large groups of individuals more vulnerable to violent narratives and extremist recruitment (Roberts-Ingleson & McCann, 2023; Rulis, 2024). This is not just a European or Western concern. Misinformation has been highlighted as a key factor in online mobilization toward violence, political controversy, and health consequences in Southeast Asia as well (Akbar et al., 2022; Angeline et al., 2020; Chang et al., 2021; Mujani & Kuipers, 2020; Quirk, 2021).

Defined by Wardle and Derakhshan as "false information shared without the intent to cause harm" (2017), misinformation, and its close cousin disinformation,[2] have become ubiquitous in online environments ranging from social media platforms and microblogging sites, to digital 'word of mouth' on chat platforms and traditional media outlets. Simply put, increased connectivity increases exposure to mis- and dis-information.

In response, organizations and governments have investigated a range of possible methods to bolster individual and society-level resilience to mis- and dis-information in recent years. Drawing upon education, media, and digital literacy scholarship, interventions have sought to help citizens identify specific narrative strategies and tactics used in misinformation and disinformation (Maertens et al., 2021; Polarization & Extremism Research & Innovation Lab, 2025). That is not to say that all interventions are alike; intervention duration and modality is varied and may range from in-depth instruction and long-term observation to shorter online modules – even presented via gamified approaches to engage users (Jeong et al., 2012). As program cost often scales with the intensity of intervention modality, specific tailored instruction and contact time between instructors and participants may balloon costs and prohibit programs from scaling effectively (Jeong et al., 2012; Lewandowsky & van der Linden, 2021; Maertens et al., 2021). By contrast, recent gamified approaches to digital literacy and mis/disinformation prevention have shown promise (Basol et al., 2020; Harjani et al., 2023; Sailer & Homner, 2020).

However, simply developing interventions is not enough. To understand *why* some interventions are effective while others are not, a strong theoretical foundation is essential. Such frameworks provide testable mechanisms by which interventions are expected to

[2] Which they define as "false information shared with the intent to cause harm."

influence individuals' attitudes, beliefs, and behaviors related to misinformation. Within this context, psychological and communication literature provide valuable guidance as to why some interventions are effective at reducing high risk online behaviors, while others may not be. In particular, the Theory of Planned Behavior (TPB) (Ajzen, 1991) outlines a specific mechanism by which changes in attitudes may translate into changes in specific subsequent behaviors – as in the case of interventions - and extending McGuire's work on inoculation (McGuire, 1961; McGuire & Papageorgis, 1961), communication research has shown to be a useful guide in shaping the content of such interventions (Compton et al., 2021; Lewandowsky & van der Linden, 2021; Roozenbeek et al., 2022; Traberg et al., 2022). By exposing participants to 'weakened' forms of misinformation and giving specific guidance on how to identify these influence behaviors, participants may become more resilient to such threats.

To date, the literature examining the impacts of gamified interventions aimed at preventing the spread of mis- and disinformation is growing, but incomplete (Basol et al., 2020; Compton et al., 2021; Harjani et al., 2023). While research has shown that mimicking native social media environments where information is consumed can improve participant ability to detect mis- and disinformation 'in the wild', few studies have explored the application of such interventions on group messaging applications – a key platform where mis- and disinformation are trafficked (Baulch et al., 2024; Harjani et al., 2023). Moreover, existing research has been largely focused on Western populations, hampering the potential generalizability of findings, or potentially obfuscating key cross-national differences. This study aims to help fill these gaps by examining the effectiveness of Gali Fakta (Indonesian for "Dig up the facts") - a gamified intervention - on intent to identify four prominent types of misinformation (fake accounts, confirmation biases, trustworthy sources, and filter bubbles) online among adults in Indonesia.[3]

We proceed first by outlining the mis- and dis-information intervention landscape, taking particular care to note impacts and examples in non-Western locales. Next, we discuss the inoculation theory and the theory of planned behavior in the context of interventions. Then

[3] We note that this list is not exhaustive, and indeed other studies have targeted further facets of mis and disinformation including impersonation, emotional content, polarization, conspiracies, discreditation, and trolling (Maertens et al., 2021).

we describe the present study, highlighting the survey sample and procedure, and summarize the empirical strategy including both ordered logistic and ordinary least squares regression approaches. Finally, we present analytical results evidencing positive, durable shifts in respondent attitudes and intentions related to online misinformation.

Literature Review

The Many Impacts of Misinformation

In the summer of 2017, a Pew Research study canvassed experts ranging from technologists to scholars and practitioners, asking "In the next 10 years, will trusted methods emerge to block false narratives and allow the most accurate information to prevail in the overall information ecosystem? Or will the quality and veracity of information online deteriorate due to the spread of unreliable, sometimes even dangerous, socially destabilizing ideas?" (J. Anderson & Rainie, 2017). While experts were evenly split at the time, the subsequent eight years have seen academic and civil society organizations document serious social, political, and human impacts of misinformation's spread (American Psychological Association, 2023; Rascoe, 2024). Though not unique to digital spaces the proliferation of misinformation in the digital age is notable for its reach and acceleration (K. Anderson, 2019). Beyond simply disseminating false or inaccurate information, misinformation erodes trust in key institutions, including governments, democratic participation, media outlets, and even scientific communities (Akbar et al., 2022; Baulch et al., 2024; Syam & Nurrahmi, 2020).

This erosion of trust can have far-reaching consequences, compounded by the fact that misinformation spreads due to well-understood, but common factors. When misinformation aligns with an individual's identity and their perceived social norms, or if it is novel or affectively charged, research shows that individuals are statistically significantly more likely to spread it (even if they do not believe it) (K. Anderson, 2019). Moreover, misinformation may focus anger or frustration at outgroups – exposing consumers to endorse or more readily accept justifications for targeted or political violence (Roberts-Ingleson & McCann, 2023; Rulis, 2024).

These social risk-factors are catalyzed when misinformation is distributed from trusted sources like friends and family (Wang et al., 2019). Paradoxically, the social and relationship capital in these relationships also provides a potential opening to safely rebut misinformation without damaging key social support systems (Syam & Nurrahmi, 2020). When individuals are trained to identify common misinformation tactics and given opportunities to practice rebutting them, they are more likely to do so, and they are more likely to view it as a normative practice (Basol et al., 2020; Harjani et al., 2023; Traberg et al., 2022).

Inoculation Theory and Prebunking Games

Inoculation theory, as introduced in the 1960s, is premised on the idea that when individuals are exposed to weak arguments against their attitudes or beliefs, they're able to develop stronger defenses against future stronger, or even more subtle attacks. This functions via 1) raising awareness that their beliefs or attitudes may become vulnerable to attack, and 2) the development of refutation strategies against weak counter arguments to their beliefs or attitudes. This is because when beliefs are threatened, people are motivated to counter-argue, which may equip them to respond more effectively. As a consequence, individuals often become more committed to their original beliefs and better equipped to resist future persuasion attempts (McGuire, 1961; McGuire & Papageorgis, 1961).

Contemporary research has further developed this framework examining nuances of different forms of threats. For instance, Compton and colleagues suggest that the type of threat matters (e.g., threat to freedom, threat to competence, etc.), and while a more generic warning about a threat may induce a weaker form of inoculation, specifying the source and specific nature of counter arguments or threats to beliefs may be more effective. Moreover, while inoculation as presented by McGuire (1961) does not necessarily differentiate individuals who are already aware or expert in a topic, Compton and colleagues (2016, 2022) suggest that inoculation may need to differ depending on topical knowledge or expertise.

This work has been applied specifically in the domain of inoculation against misinformation or fake news. Roozenbeek and colleagues (2022) show that when individuals are exposed to weaker versions of common misinformation tactics, such as fake experts (e.g., a seemingly qualified individual in a lab coat making unsupported scientific claims) or

appeals to emotion (e.g., using emotionally charged language or images to bypass thoughtful consideration), and they are informed as to how these tactics are being used to manipulate them, the spread of misinformation is reduced. This is referred to as the idea of prebunking, or giving individuals tools to identify and resist misinformation before they are exposed to specific false claims (Braddock, 2022; Carthy & Sarma, 2021; Roozenbeek et al., 2022). Work motivated by this framework has leveraged strategies such as short videos, brief explanatory documents, and interactive games to inoculate people against common misinformation tactics (Piltch-Loeb et al., 2022; Schumann & Barton, 2024).

Theory of Planned Behavior

While prebunking interventions aim to equip individuals with the skills to resist misinformation, understanding how these interventions influence behavior requires a robust theoretical framework. The Theory of Planned Behavior (TPB), developed by Ajzen (1991) provides such a framework, tracing the determinants of behavioral intentions and presents a model linking those intentions with specific behavioral outcomes. Briefly, the theory posits that human behavior is best predicted by specific intention to perform that behavior, and behavioral intentions are shaped by three factors:

1) Attitudes toward the behavior (how one feels about performing the action),
2) Subjective norms (perceived social pressure around the behavior), and
3) Perceived behavioral control (self-efficacy: the extent to which individuals believe they can successfully perform the behavior).

Applied to the context of mis- and dis-information, TPB suggests that interventions should aim to positively influence attitudes toward identifying misinformation, foster subjective norms that support critical evaluation of online content, and enhance individual's perceived control over identifying and avoiding misinformation. The prebunking games, based on inoculation theory, have shown promise in combating misinformation by exposing individuals to small doses of false information in controlled settings (Basol et al., 2020, 2021; Compton et al., 2021).

However, these interventions, typically designed for Western audiences, may not be culturally relevant elsewhere (Badrinathan & Chauchard, 2024; Henrich et al., 2010). For instance, Harjani and colleagues (2023) developed a media literacy game for participants in north India but found it did not improve their ability to evaluate or reduce their willingness to share misinformation. The authors suggest future interventions would be more effective if adapted with input from local researchers and universities. They also highlight the need to consider the limited digital literacy of rural audiences, which may impact the effectiveness of these interventions (Harjani et al., 2023).

The Indonesian Context

In this study, we focus on Indonesia, the fourth-most populous country in the world. Recently, Indonesia has experienced a notable increase in digital engagement, with 77% of its population currently online (Sharon, 2024; Zhulfakar, 2024). Projections indicate that this figure could rise to 90% by 2025 (Kemp, 2023; Nurhayati-Wolff, 2023). However, this boost in online connectivity has also heightened exposure to misinformation. Following the re-election of Indonesia's president Joko Widodo in 2019, supporters of the opposition candidate alleged widespread electoral fraud - a claim that was amplified through coordinated disinformation campaigns conducted primarily on encrypted platforms like WhatsApp and Telegram (Theisen et al., 2021). These channels were used to circulate fabricated images, conspiracy theories about foreign interference, and false claims regarding violence against opposition supporters. The disinformation helped ignite unrest that culminated in large-scale protests in Jakarta in May 2019, leading to violent clashes with security forces, resulting in at least six deaths and over 200 injuries (BBC News, 2019). Additional research by Rahmawan and colleagues (2024) analyzed WhatsApp-related hoaxes verified by Indonesian fact-checking organization MAFINDO between 2015 and 2020, finding that politically charged misinformation was among the most frequently shared, often spreading in intimate settings like private and family group chats.

Given WhatsApp's widespread use in Indonesia and its vulnerability to such misinformation (Baulch et al., 2024), we investigate Gali Fakta, a gamified intervention designed to improve media literacy skills in a simulated WhatsApp environment to address

this need. Previous research conducted by Facciani and colleagues (2024) using an earlier iteration of Gali Fakta as a gamified media literacy intervention, indicated that those who played the game demonstrated discriminatory skepticism towards false news headlines and reduced the likelihood of sharing them.[4]

Hypotheses

Given this theoretical and empirical context, we present three sets of hypotheses. First, per inoculation theory we hypothesize that after being exposed to the intervention, participants will demonstrate growth in the three antecedent areas described by the theory of planned behavior (attitudes, norms, self-efficacy), as well as behavioral intentions themselves. More formally:

1a) After participating in Gali Fakta, participants will report more positive behavioral attitudes around detecting and preventing misinformation.
1b) After participating in Gali Fakta, participants will report stronger subjective norms around detecting and preventing misinformation.
1c) After participating in Gali Fakta, participants will report higher self-efficacy around detecting and preventing misinformation.
1d) After participating in Gali Fakta, participants will report greater behavioral intentions to detect and prevent misinformation.

Second, and consistent with the Theory of Planned Behavior (Ajzen, 1991), we expect that:

2) Respondents who report stronger behavioral attitudes, subjective norms, and self-efficacy around detecting and preventing misinformation will report greater behavioral intentions to detect and prevent misinformation.

[4] By discriminatory skepticism, the authors describe skepticism toward fake news headlines, but no generalized skepticism toward factual news headlines.

Third (3) and pragmatically speaking, we expect that participants will tend to improve their media literacy skills as evidenced by higher post-instruction (vs. pre-instruction) gameplay scores.

After participating in Gali Fakta, participants will score higher on media literacy than before participating in the intervention.

Data and Methods

To test these hypotheses, we leverage anonymized participant data collected by Moonshot between July and September 2024. Players were recruited through Bilendi & Respondi (a research firm) and demographic quotas were set to recruit a sample representative of the Indonesian population. In particular, the firm ensured representation of age brackets, and gender by age bracket.[5] All respondents affirmed that they were located in Indonesia and all elected to play the game in Indonesian (vs. English).

Of the (n = 650) players of the game 504 (77.5%) had complete gameplay data and played the game within feasible, minimum time constraints.[6] This is congruent with research that has shown that an estimated 35% of participants tend to respond inattentively or otherwise carelessly on computer-administered surveys: which, research also has shown, can obscure the genuine effects of an intervention (Maniaci & Rogge, 2014; Oppenheimer et al., 2009).

Methodology

Upon recruitment by Bilendi & Respondi, respondents were sent a URL with the initial screening survey (to achieve demographic quotas) - after which they were redirected to the gamified intervention designed to improve media literacy within players' information environment with the aim of boosting prebunking effectiveness. The game was designed in conjunction with Indonesian game designers and disinformation experts to refine the

[5] Defined age brackets were (1) 18-24, (2) 25-34, (3) 35-44, (4) 45-54, (5) 55-64, and (6) 65+

[6] Minimum acceptable time limit was ≥ 50 seconds at the completion of the pretest. This comported with the game designers who—through first-hand experience—suggested that the game should take approximately one minute to play through the pretest items. See discussion for a more detailed treatment of the attrition.

aesthetics, clarity, and overall experience. In Indonesia, existing laws prohibit spreading false information or glorifying misinformation (Prahassacitta & Harkrisnowo, 2021), so the game aimed to empower users to protect their friends and family from hoaxes online. The final game, called "Gali Fakta" was developed through collaboration between Moonshot, the University of Notre Dame, and IREX, and draws on IREX's Learn to Discern Program (IREX, n.d.).[7]

The game itself takes approximately five minutes to play and simulates a group chat where players are prompted to protect their friends and family (i.e., those with whom they frequently speak using messaging applications) from hoaxes. Here they choose from three content areas—health, news, or finance—and encounter lessons on identifying four types of misinformation: fake accounts, confirmation bias, trustworthy sources, and algorithmically generated filter bubbles (collectively referred to as misinformation moving forward). Throughout the game, players engage in four scenarios to learn about misinformation tactics, with the script reviewed by Indonesian experts. In sum, the game aimed to inoculate users by warning them about potential misinformation (*specific threat*) from loved ones (*specific sources*) and teaching them how to detect it through simulated conversations.

Before completing the game, participants were asked about their age and gender. They were then directed to play and complete the game. Once completed, they were asked to report their attitudes via a retrospective pre-post format.[8] They were then contacted two weeks after their participation and they were again asked about their *self-assessed* attitudes, norms, self-efficacy, and behavioral intentions around identifying fake accounts, confirmation bias,

[7] Specific materials of the intervention are available upon reasonable request.

[8] A pre-post retrospective format was used, whereby participants were asked to reflect on their attitudes *before* playing the game *after* completing it. This approach, while not without limitations (expanded upon in the Discussion section), is valuable in detecting changes in potentially undesirable behaviors, as well as when attitudes are fundamentally linked with knowledge. Retrospective pre-post design guards against potential response shift bias – the extent to which respondents' pre-post responses differ because of a shift in their understanding of themselves or an issue, rather than as a consequence of an intervention (Geldhof et al., 2018; Howard & Dailey, 1979; Moore & Tananis, 2009). This is germane to the present study, given (for example) that—if a respondent does not realize that they are relatively unaware of online risks—they are likely to over-estimate their pre-test awareness (Geldhof et al., 2018; Pratt et al., 2000).

trustworthy sources, and filter bubbles online.[9] In total, respondents provided these attitudes at three time points.

Survey Instrument

Below, we note the specific measurement and operationalization of the key measures we report in this study. Beginning with the four core concepts of the Theory of Planned Behavior, each concept was measured using one 6-level (Strongly Disagree = 1, Strongly Agree = 6) Likert-type question to assess their agreement with the following statements.

> **Attitudes**: "I want to be able to identify fake accounts, confirmation bias, trustworthy sources, and filter bubbles when I encounter them online."
> **Norms**: "It is normal to identify fake accounts, confirmation bias, trustworthy sources, and filter bubbles when I encounter them online."
> **Self-efficacy**: "I feel able to identify fake accounts, confirmation bias, trustworthy sources, and filter bubbles when I encounter them online."
> **Intentions**: "I will identify fake accounts, confirmation bias, trustworthy sources, and filter bubbles when I encounter them online."

In addition to these attitudinal items, participant performance in the Gali Fakta learning intervention was recorded before the instructional portion of the intervention, and again after they had completed the instructional scenarios. The four test scenarios presented after completing the instructional portion covered substantively similar topics and techniques as in the first half, however the specific details and wording were distinct to avoid a training bias or inflated scores due to player recall. As noted above, players were tested on four domains (i.e., fake accounts, filter bubbles, trustworthy sources, and confirmation bias) at both time points (before and after the instructional element of the intervention); their responses were scored as correct or incorrect, and their total score was recorded (0-4) at point

[9] A two-week period was selected to balance respondent retention (i.e., avoiding attrition), and assess the potential longitudinal impact. As has been done by Maertens and colleagues (2021), future work should consider additional time windows to assess potential decay/persistence of intervention impacts.

points. In particular, we highlight the initial participant score and the difference between the pre-training and post-training score (to account for potentially different starting points).

Respondents also provided demographic information including gender, age, education level, and typical time spent online. *Gender* was reported as a self-nominated category including "Female", "Male", and "Prefer not to say". Respondent *Age* was reported in years, and *Education Level* was obtained by asking for the highest level of education achieved (categorized in an ordinal measure including Primary (Grades 1-6), Junior Secondary (Grades 7-9), Senior Secondary (Grades 10-12), College or University, and Post-College Graduate Degree). Finally, participants were asked how much time they spend online on average. Ordinal response categories included 1) 0-Less than one hour per day; 2) 1-2 hours per day; 3) 3-4 hours per day; 4) 5-6 hours per day; and 5) 7+ hours per day.

Analytic Strategy

To address the research questions and specific hypotheses above, we proceed in three main stages: first, by presenting descriptive statistics of the sample; next by estimating the ordered logit regression models; and finally by estimating the ordinary least squares regression models. All statistical analyses were conducted in R version 4.5.0 using tidyverse, *polycor*, *likert*, *lmtest*, *sandwich*, *jtools*, and *Hmisc* packages.

Following the descriptive examination, we estimate a series of ordered logistic regression models for each of the four components of the theory of planned behavior. Since our primary dependent variables are single Likert items, which are inherently ordinal in nature, ordered logistic regression (OLOGIT) is the most appropriate statistical approach and accounts for the ordered categories of the outcome while relaxing the assumption of equal spacing between categories, which would be violated by standard linear regression. Specifically, these models allow us to examine whether these changes across waves were statistically reliable, and to what extent they may have depended on relevant covariates. Additionally, we assess the empirical robustness of the models – contrasting the model predicted class against a naïve prior. Finally, we report ordinary least squares models

considering the application of the theory of planned behavior at Wave 3 (the 2-week follow-up).[10]

Results

Descriptive Statistics

Beginning with univariate examination, Table 1 presents descriptive statistics for the respondents and Figures 1 through 4 illustrate the four TPB measures over time (pre-test, post-test, and follow-up).[11]

Considering the demographic characteristics of the sample, the participants in this study were almost evenly split between men and women, and the average age was in their late 30s. Further this was an educated sample with over 90% of respondents having completed work at the senior secondary, college, or university levels.

Next, information on typical time online was included with the expectation that individuals who spend more time online could be more aware of or have already experienced online mis- and disinformation tactics. Most participants in this study reported spending over 3 hours a day online, with less than 12% reporting zero to two hours per day.

Table 1: Sample Descriptive Statistics

Characteristic	**N = 504[1]**
Pretest Game Score	
0	96/504 (19%)
1	176/504 (35%)
2	126/504 (25%)
3	82/504 (16%)
4	24/504 (4.8%)
Pre-Post Score Change	0.55 (1.15)
Online Time	
0 - Less than one hour a day	1/504 (0.2%)
1-2 hours a day	56/504 (11%)
3-4 hours a day	170/504 (33%)

[10] Hereafter, wave 1, 2, and 3 are used to refer to data collected at Pre-test, Post-test, and Follow-up, respectively.
[11] Estimated heterogeneous correlation matrix for all indicators is available upon request.

5-6 hours a day	145/504 (29%)
7+ hours a day	132/504 (27%)
Gender	
Female	247/504 (51%)
Male	257/504 (49%)
Age	39 (13)
Highest Education Completed	
Primary (Grades 1-6)	5/504 (1.0%)
Junior Secondary (Grades 7-9)	13/504 (2.6%)
Senior Secondary (Grades 10-12)	232/504 (46%)
College or University	229/504 (45%)
Post-college Graduate Degree	25/504 (5.0%)

[1]n/N (%); Mean (SD)

Finally, we note respondent performance in the game-related tasks. Prior to engaging in the instructional element of the game, the typical respondent was able to correctly provide 1 correct answer (out of 4). That is, at baseline, respondents were quite poor at detecting fake accounts, identifying the influence of filter bubbles, highlighting trustworthy sources, and uncovering confirmation bias. After the training portion, respondents answered a median of 2 (of 4) questions correctly suggesting limited improvement, with an average score increase of 0.55 correct items.[12]

Next, considering Figures 1 through 4, we find that behavioral attitudes, subjective norms, and behavioral intentions around detecting misinformation tended to increase over the period of the study. By contrast, respondent self-efficacy improved between the pre-test and post-test levels, however they returned to baseline in the two-week follow-up period.

[12] This difference is statistically significant at $p < 0.001$ using a Wilcox Signed-Rank Test ($V = 70714.5$) (Rosner et al., 2006).

Figure 1: Behavioral Attitudes by Wave

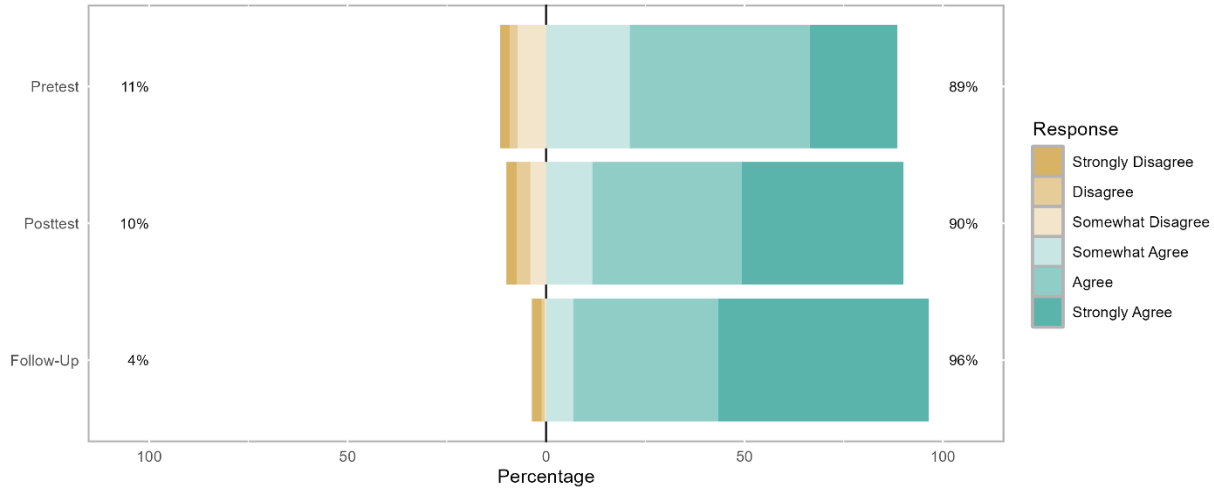


Figure 2: Subjective Norms by Wave

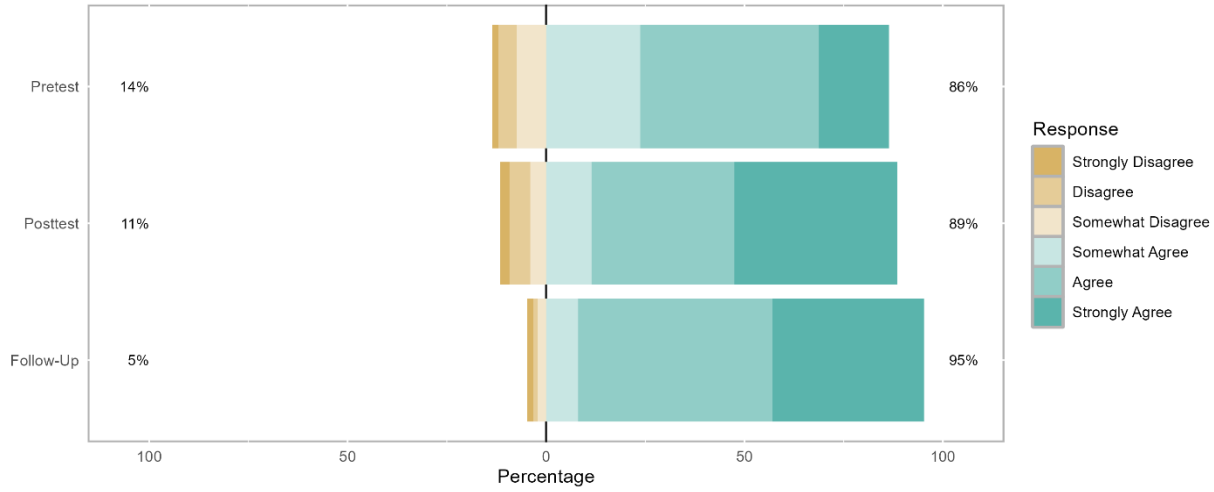


Figure 3: Self-Efficacy by Wave

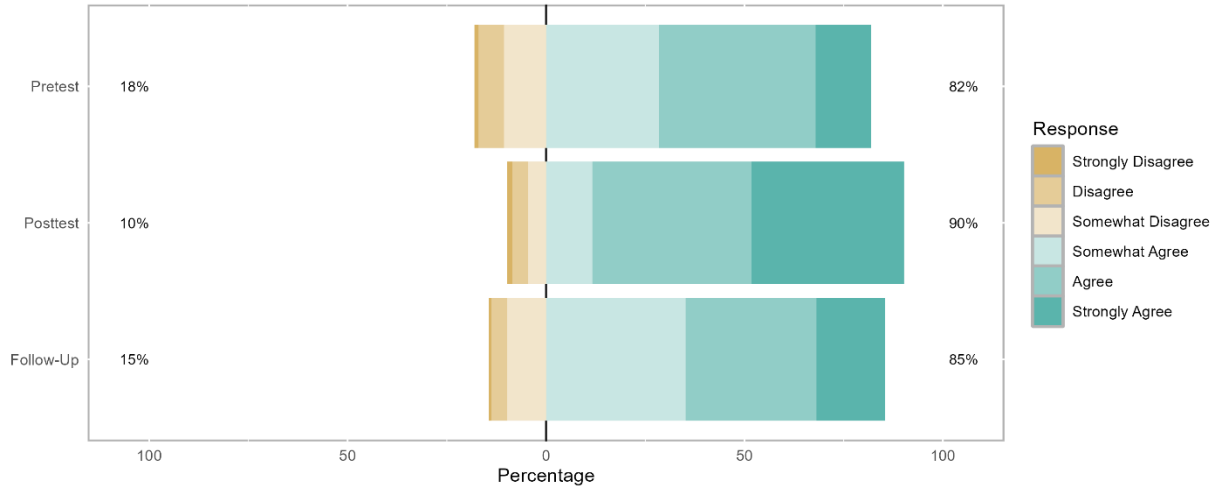
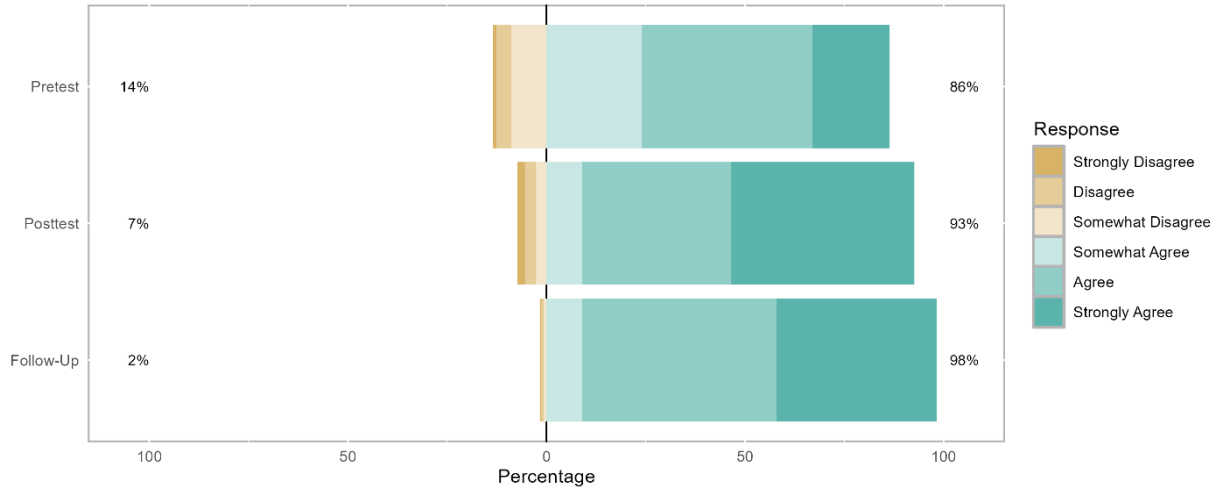


Figure 4: Behavioral Intentions by Wave



TPB – Change over Time

Moving on to the specific hypothesis tests, Table 2 presents the results of the ordinal logistic (OLOGIT) regressions examining changes in each of the TPB concepts over time, empirically testing H1a-d. Since each of the TPB concepts is measured using a single Likert item, OLOGIT is an appropriate estimator that accommodates the structure of the dependent variables (e.g., it does not assume such variables to be normally distributed, and—hence—provides reliable estimates, regardless of whether such variables are normally distributed). As follows, we highlight the statistically significant findings for each of the four TPB concepts as they changed across the three periods. For all models, standard errors are clustered on the respondent ID to avoid over-stating precision and yield within-participant interpretations to the estimates.[13]

[13] Please note that all coefficients in Table 2 are reported as changes in the log-odds of being in a higher order category; specific interpretations of point estimates (though potentially misleading outside of this sample) may be obtained by exponentiating the coefficient and interpreting it instead as changes in the odds ratio of being in a higher ordered category. For the purposes of this study, we merely note when specific covariates were associated with statistically significant increases/decreases in the log-odds to avoid potentially misleading readers.

Table 2: Ordered Logistic Regression Coefficient Table

	Behavioral Attitude			**Subjective Norms**			**Self-Efficacy**			**Behavioral Intention**		
Variable	Coef.	SE	Sig	Coef.	SE	Sig	Coef.	SE	Sig	Coef.	SE	Sig
Wave:												
Post-Test	0.701	0.110	***	0.898	0.116	***	1.224	0.112	***	1.288	0.123	***
Follow-Up	1.341	0.107	***	1.059	0.101	***	0.054	0.093		1.251	0.104	***
Initial Score	0.059	0.067		0.011	0.066		0.002	0.067		0.039	0.073	
Score Change	0.069	0.064		0.054	0.059		0.049	0.063		0.079	0.064	
Online Time	0.260	0.061	***	0.199	0.059	***	0.296	0.060	***	0.320	0.058	***
Education	0.142	0.090		0.142	0.086	†	0.180	0.094	†	0.165	0.092	†
1|2	-1.629	0.348		-2.294	0.362		-2.607	0.393		-2.143	0.398	
2|3	-0.968	0.323		-1.063	0.321		-0.779	0.316		-0.845	0.333	
3|4	-0.298	0.319		-0.408	0.314		0.217	0.307		-0.001	0.319	
4|5	0.862	0.314		0.700	0.310		1.680	0.307		1.292	0.314	
5|6	2.773	0.322		2.697	0.318		3.466	0.317		3.363	0.325	
AIC	3787.2			3942.9			4237.1			3685.3		

Note: *** $p<0.001$; ** $p<0.01$; * $p<0.05$; † $p<0.10$; Coefficients are un-transformed log-odds.
Standard errors clustered on respondent ID. Wave coefficients estimated in reference to the pre-test value.

First, considering behavioral attitudes, consistent with Figure 1, we find that across the three waves respondents had greater odds of being in a higher response category in the Post-instruction wave relative to the Pre-instruction period (Coef.= 0.701, SE = 0.110, $p < 0.001$). Moreover, when rotating the reference category (to the Post-Test score), we find that behavioral attitudes in Wave 3 are higher (Coef.= 0.640, SE = 0.110, $p < 0.001$). Additionally, we find that individuals who reported spending more time online on average indicated greater behavioral attitudes (Coef.= 0.260, SE = 0.061, $p < 0.001$). We did not find evidence of a statistically significant relationship between initial performance on the game or change in performance in the game (i.e., post-instruction), with education level or responses to the survey items.

Next, we found a similar pattern for subjective norms, with respondents reporting higher levels at post-test compared to pre-test (Coef.= 0.898, SE = 0.116, $p < 0.001$). Further, follow-up scores were greater than the pre-test item (Coef.= 1.059, SE = 0.101, $p < 0.001$). Likewise, we find that individuals who reported spending more time online on average reported greater subjective norms around identifying misinformation (Coef.= 0.199, SE =

0.059, $p < 0.001$). When rotating the reference category to the post-test value, we do not find evidence of a statistically significant difference with the follow-up test (Coef.= 0.162, SE = 0.112, $p = 0.150$), indicating that (with one exception) results held at the third/final follow-up reference point.

Regarding the aforementioned exception we find a divergent pattern of results only for participant self-efficacy related to identifying potential misinformation. Relative to the baseline pre-test score, on average respondents reported greater self-efficacy at the post-test section of the retrospective pre-post (Coef.= 1.224, SE = 0.112, $p < 0.001$), however we observed no evidence of a difference between pre-test and follow-up scores. Indeed, when rotating the reference category to the post-test, self-efficacy at the follow-up had decreased (Coef.= -1.170, SE = 0.112, $p < 0.001$). In this model, we continued to find evidence that time online was positively associated with our outcomes such that those who spent more time online reported higher self-efficacy (Coef.= 0.296, SE = 0.060, $p < 0.001$).

Finally, we find evidence that respondent behavioral intent to identify misinformation increased across waves. That is, relative to the pre-test, respondents indicated greater behavioral intent at post-test (Coef.= 1.288, SE = 0.123, $p < 0.001$), as well as in the two-week follow-up (Coef.= 1.251, SE = 0.103, $p < 0.001$). We did not, however, find evidence that behavioral intent continued to increase following the post-test period (Coef.= -0.037, SE = 0.101, $p = 0.716$). As in the previous models, we observed that online time remained positively associated with the outcome (Coef.= 0.320, SE = 0.058, $p < 0.001$). In sum, we observe that though there was no evidence that initial task performance or change in game task performance had a statistically significant impact on the behavioral attitude, norms, self-efficacy, and behavioral intent, respondents reported statistically significant changes over time on these attitudes, and in the case of all but self-efficacy, these upward shifts persisted at the two-week follow-up.

To assess the empirical robustness of these models, Tables 3-6 present the confusion matrices for each of the four components of TPB (attitude, norms, self-efficacy, and intent). Confusion matrices are a method of illustrating the difference between predicted and observed values in models estimating categorical outcomes. Here we determined the model-predicted class based on the maximum predicted probability from the models in Table 2. Since each of

the TPB measures is an ordinal outcome, we illustrate the overall accuracy of the model predictions contrasted with the No Information Rate (NIR) (Altman & Bland, 1994; Kuhn, 2008)[14]. In the case of behavioral intentions, attitude, and perceived norms, the predicted values represent an improvement on the NIR, however we note that the self-efficacy model did not yield a statistically significant increase in prediction accuracy.

Table 3: Confusion Matrix – Behavioral Attitudes

		Predicted					
		1	**2**	**3**	**4**	**5**	**6**
Observed	**1**	0	0	0	0	23	13
	2	0	0	0	0	26	6
	3	0	0	0	0	52	6
	4	0	0	0	0	153	48
	5	0	0	0	0	***366***	237
	6	0	0	0	0	232	***350***

Note: Bolded and Italicized values correctly predicted**.** Predicted and observed values include performance on all three waves.
No Information Rate = 0.399
Accuracy = 0.474

[14] Though class-wise precision and specificity may also be reported for such ordinal outcome models, these are calculated by cell and thus in lieu of reporting class-wise diagnostic metrics, we choose to focus on the NIR and global accuracy measures. The NIR represents the baseline accuracy that would be achieved by simply guessing the most frequent category.

Table 4: Confusion Matrix – Subjective Norms

		Predicted					
		1	**2**	**3**	**4**	**5**	**6**
Observed	**1**	0	0	0	0	19	6
	2	0	0	0	0	48	9
	3	0	0	0	0	63	4
	4	0	0	0	0	189	26
	5	0	0	0	0	***525***	128
	6	0	0	0	0	329	***166***

Note: Bolded and Italicized values correctly predicted**.** Predicted and observed values include performance on all three waves.
No Information Rate = 0.432
Accuracy = 0.457

Table 5: Confusion Matrix – Self-Efficacy

		Predicted					
		1	**2**	**3**	**4**	**5**	**6**
Observed	**1**	0	0	0	3	9	3
	2	0	0	0	17	51	5
	3	0	0	0	29	90	6
	4	0	0	0	***68***	295	19
	5	0	0	0	69	***417***	76
	6	0	0	0	17	233	***105***

Note: Bolded and Italicized values correctly predicted**.** Predicted and observed values include performance on all three waves.
No Information Rate = 0.372
Accuracy = 0.390

Table 6: Confusion Matrix – Behavioral Intentions

		Predicted					
		1	**2**	**3**	**4**	**5**	**6**
Observed	**1**	0	0	0	0	11	4
	2	0	0	0	0	32	6
	3	0	0	0	0	53	9
	4	0	0	0	0	168	43
	5	0	0	0	0	***432***	209
	6	0	0	0	0	262	***283***

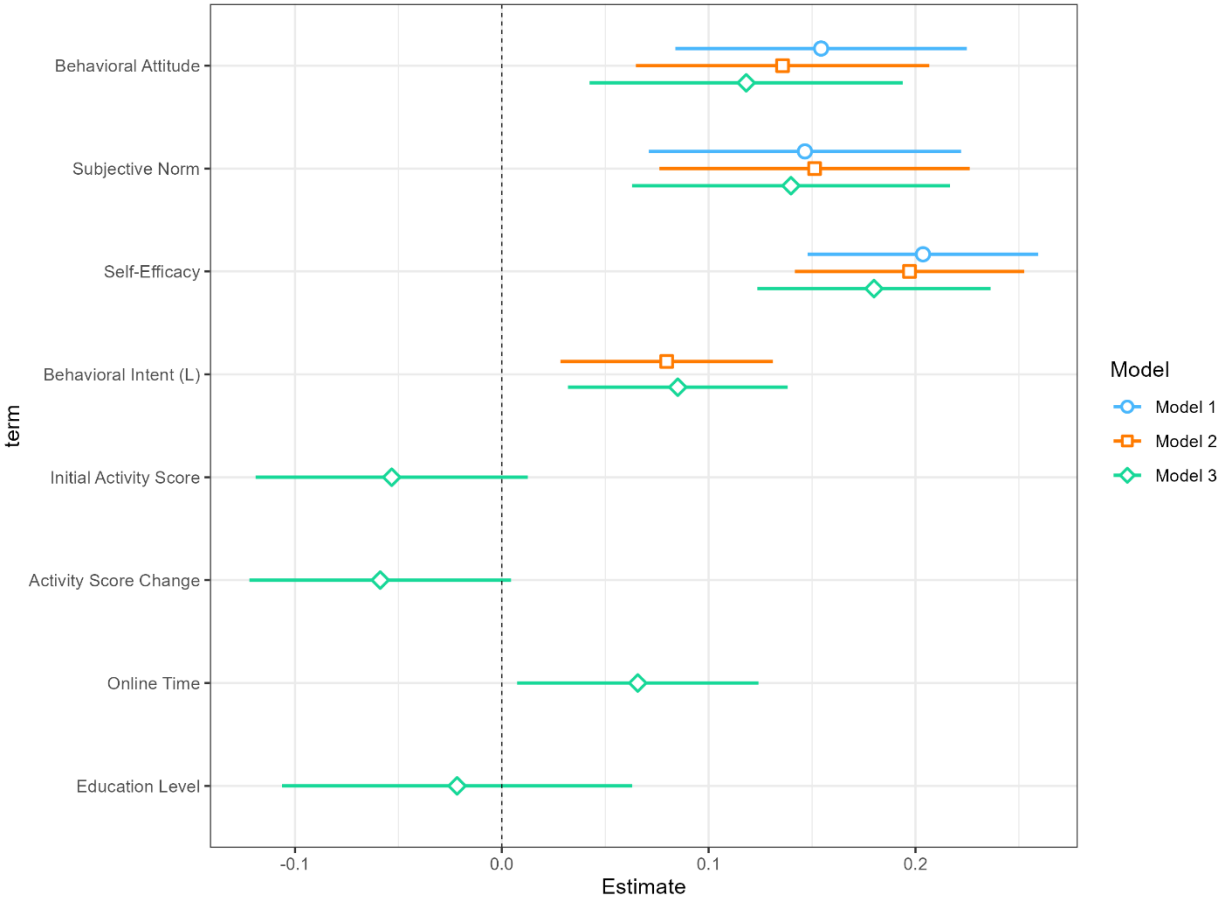
Note: Bolded and Italicized values correctly predicted**.** Predicted and observed values include performance on all three waves.
No Information Rate = 0.424
Accuracy = 0.473

TPB – Antecedents of Behavioral Intent

In considering the antecedents of behavioral intent, we find support for Hypothesis 2 (i.e., stronger behavioral attitudes, subjective norms, and self-efficacy were positively associated with participants' behavioral intentions). Figure 5 presents the results of the ordinary least squares regression of Wave 3 behavioral intention on Wave 3 attitudes, norms, and self-efficacy, Wave 2 behavioral intention, respondent baseline game performance and change, their time online, and their self-reported level of education. In this figure, the vertical line at 0.0 represents the null hypothesis (i.e., insufficient evidence of a relationship), and point estimates include their respective 95% confidence intervals.[15] As in the models presented above, standard errors are clustered on the respondent to account for meaningful within-unit variation.

[15] Where these 'whiskers' intersect the threshold, we do not have evidence to suggest that the impact/relationship is distinct from zero; conversely, where whiskers do not intersect zero, we conclude that the estimate is statistically significant.

Figure 5: Coefficient Plot for TPB on Wave 3 Behavioral Intent



Looking first at Model 1, we find initial support for H2 – that is, contemporary behavioral attitudes, subjective norms, and self-efficacy around detecting online misinformation are each positively associated with the concomitant behavioral intention when controlling for the other indicators. Moving to Model 2, we find that these relationships remain positive and statistically significant even when controlling for the 2-week lagged measure of behavioral intentions. Unsurprisingly – we find that lagged behavioral intention (Behavioral Intent (L)) is also positively related to initial, post-instruction behavioral intent for identifying misinformation. Here, the coefficient estimate is smaller in magnitude than each of the other TPB items (point estimate of 0.089 vs. 0.118, 0.138, and 0.186 for attitude, norm, and self-efficacy respectively).

Finally, when incorporating initial game performance, activity score change, online time, and education level, we find that support for H2 is robust to these potential confounders.

Indeed, the coefficient estimates for the three TPB antecedents are statistically indistinguishable across models. We do note though that among the four control variables, only the ordinal measure of online time was positively related to intent to identify online misinformation. That is, individuals who reported spending more time online also indicated greater intent to identify misinformation when they encounter it online.

Stepping back to consider relative fit for the three models, we note that all three explained over 20% of the variation in the dependent variable (W3 Behavioral Intentions). Contrasting the adjusted R^2 values, Model 1 explained 23.0% of the variation, Model 2 explained 24.6% of the variation, and Model 3 explained 25.4% of the variation: a large impact by convention (Cohen, 2013). Moreover, Model 2 was a statistically significant improvement over Model 1 (that is, adding the lagged behavioral intention statistically significantly improved fit; $F = 11.16$, $p<0.001$), and Model 3 - collectively adding original task performance, task performance change, online time, and education level was a statistically significant improvement over Model 2 ($F = 2.40$, $p < 0.05$). The F-test results indicate that adding the lagged behavioral intention and the demographic controls improved the model's explanatory power for Model 1-2 and 2-3 respectively.

Discussion

The results of these analyses suggest that the Theory of Planned Behavior (Ajzen, 1991) provides a valuable framework for assessing the potential impacts of mis- and dis-information interventions – particularly when measuring behavioral outcomes directly is impossible or otherwise infeasible. Our findings indicate that Gali Fakta players reported increased desire, perceived norms, and intent to identify online misinformation (including fake accounts, confirmation bias, trusted sources, and algorithmic filter bubbles) following a brief WhatsApp styled intervention. Through this work, we contribute to the growing body of literature on the potential benefits of mis- and dis-information prevention interventions – consistent with extant work on inoculation and media literacy more generally (Basol et al., 2021; Braddock, 2022; Carthy & Sarma, 2021; Facciani et al., 2024; Jeong et al., 2012).

Drawing on the results presented above, we provide evidence of support for H1a, H1b, and H1d[16], as for each of these 3 outcomes, respondents retained the increases following their participation up to two weeks later (or further improved from Wave 2). By contrast (see H1c), while respondents reported an initial increase in their self-assessed capacity to identify fake accounts at Wave 2, they appeared to return to their baseline two weeks later. In Wave 3, the modal response category indicated that respondents "Somewhat Agree" with the statement that they feel able to identify online misinformation. That is, the mean respondent did not have a particularly pessimistic outlook, but rather they did not express statistically significant improvements with respect to confidence in their self-efficacy. This post-training dip, while not desirable, is well-documented in educational and training intervention settings (Albee et al., 2019; J. J. Chen & Krieger, 2023), suggesting a potential role for continued education, to maintain individuals' sense of self-efficacy with respect to media literacy.

Turning to H2, our results show support for the Theory of Planned Behavior in these data. When accounting for lagged values of behavioral intentions, respondent attitudes, norms, and self-efficacy demonstrate a positive association with behavioral intent at Wave 3. This further supports the body of research on TPB (Brehmer, 2023; Cooke et al., 2016), joining Pundir and colleagues (2021), Alwreikat (2022), and Chen and Fu (2022) to highlight the utility of TPB as a framework in the mis- and dis-information prevention space.

Interestingly, we did not find evidence of a statistically significant relationship between participant education and game performance with behavioral intentions. In terms of education, this could suggest that shared variance for educational attainment and behavioral intentions may have been subsumed by other theoretical or control variables, or that the training itself was similarly impactful regardless of formal education. Examining the possibility of multicollinearity, we do not find affirmative evidence of this first possibility. This is also consistent with the null finding of game performance on behavioral intent; we posit that this intervention may have been most effective at raising participant awareness

[16] H1a) After participating in Gali Fakta, participants will report more positive behavioral attitudes around detecting and preventing misinformation.
H1b) After participating in Gali Fakta, participants will report stronger subjective norms around detecting and preventing misinformation.
H1d) After participating in Gali Fakta, participants will report greater behavioral intentions to detect and prevent misinformation.

around misinformation, rather than necessarily teaching specific skills. We interpret this favorably: that participants tended to experience beneficial outcomes regarding their behavioral intentions with respect to media literacy, regardless of whether they were either "good" at playing the game or their level of prior education. Subsequent research should continue to investigate these specific mechanisms and the alignment between program goals and these observed outcomes.

Finally, the analyses above provide evidence in support of H3 – that players would improve their ability to identify fake accounts, confirmation biases, trustworthy sources, and filter bubbles over the course of the game. As noted above, participants improved on their pre-instruction score by 0.55 correct answers (out of a total of four). This shifted the median score from one to two correct answers, suggesting that there remains an opportunity to improve the game outcomes despite being quite promising for a brief intervention.

Together, the findings of this study contribute to the ongoing body of scholarship applying inoculation or prebunking-based interventions in the mis- and dis-information space. We also note the importance of testing the theoretical models and hypotheses outside of traditionally represented populations in the psychological and security literature (Van der Wal, 2015). Though we test the hypotheses generated by the TPB here, we note additional theoretical models bear consideration including the 3N (Needs, Narratives, and Networks) model, that provides a plausible framework for the translation of specific misinformation to targeted and political violence (Kruglanski et al., 2022; Kruglanski, 2019).

The increasing capacity of locally based survey firms and partners to secure high-quality samples opens further opportunities to examine the applicability of interventions beyond traditionally represented populations. This is a priority, especially in the context of misinformation research. Speaking specifically to the case of Indonesia, rapid population and digital growth has highlighted the importance of identifying methods of mitigating the harms of online mis- and disinformation. As others have evidenced (Theisen et al., 2021), failing to act has the potential to result in serious educational, and political consequences (Baulch et al., 2024; McDonnell & MacKinnon, 2020; Syam & Nurrahmi, 2020), or indeed susceptibility to violent narratives (Carthy & Sarma, 2021; Kruglanski et al., 2022).

Despite these promising findings, we note four key limitations to this study. Firstly, we acknowledge an important distinction between behavioral intentions and subsequent behavioral outcomes. While the theory of planned behavior is a well-established framework for understanding how attitudes may translate to behaviors, research has shown that these intentions and attitudes do not perfectly predict later behaviors even when highly specific (Cooke et al., 2016; Kroke & Ruthig, 2024). Though we did not explicitly assess respondent use of diligent online practices in the present study due to sample contact limitations, we strongly encourage future studies to incorporate objective or behavioral measures, connecting the final links of the TPB model.

Secondly, we recognize the potential biases associated with self-report data in the context of socially desirable outcomes. This is twofold: respondents contrasted their attitudes before/after the game at a single timepoint, a design choice that could lead compliant respondents to overstate potential shifts in attitudes. Further, participants may have generally overreported their intentions or perceived use of diligent online practices to illustrate prosocial behaviors and present themselves in a positive light – potentially leading to overstated coefficient estimates (or false positives). To accommodate this, we estimated all empirical models with clustered standard errors on the respondent. Moving forward, interventions and practitioners should carefully weigh the tradeoff between sensitivity to false positives and false negatives based on programmatic priorities and instrument design choices.

Third, we acknowledge the implications of differential attrition in the analytic sample relative to the original theoretical stratified sample. Specifically, this was driven by respondent 'speeding' (i.e., inattentive response patterns). Young female respondents (18-24) were statistically significantly more likely to be excluded due to speeding than older female respondents (65+) and two categories of older male respondents (55-64, and 65+). We also observed greater exclusion among male respondents in the 18-24 category relative to older male respondents (55-64 and 65+).

In sum, that the youngest age bracket was, in some respects, more prone to speeding is theoretically unremarkable and unsurprising. Nevertheless, these observations warrant further caution when interpreting the findings. Specifically, we note that in addition to age, unobserved correlates related to both inattentive response patterns and behaviors around

misinformation could not be measured. Consequently, these findings may not generalize to younger populations as a whole, but (for example) to those who demonstrate 'slower' and more deliberate cognitive processing (Tversky & Kahneman, 1986). Arguably, attentive youths reflect a 'lower risk' subset of the population, and so to understand the impacts among young adults, designers and interventionists should continue to thoroughly engage with methods of better capturing participant attention and engagement. Furthermore, it suggests that interventions intending to target a younger (e.g., 18 – 24 year old) demographic should expect above-average attrition from this age group and should increase recruitment quotas for that demographic accordingly.

Finally, we highlight that the present intervention did not include a control group and thus shifts in attitudes and behavioral intentions over time cannot be categorically attributed to the intervention/game. For instance, drawing respondent attention to the issue of mis- and dis-information (by merely surveying attitudes around it) may encourage diligence when online. Likewise, evolving cultural awareness around the risks of mis- and disinformation online may have increased over the period of study net the influence of the short game. This is an important issue to balance however, as unnecessarily withholding benefits from participants and beneficiaries is a salient ethical concern – especially when the intervention is believed to improve participant outcomes (Huston & Peterson, 2001). To apply a known practice in other settings (Batomen & Benmarhnia, 2024; Bernal et al., 2017), future interventions should consider using a staggered rollout (a so-called "waitlist control" design) to facilitate experimental or quasi-experimental identification of program impacts.

Taken together, we provide evidence that brief gamified approaches to mis- and disinformation prevention have the potential to yield promising results. Further, we highlight the importance of culturally sensitive approaches to intervention; when the game is customized to reflect common sources of potential mis- and dis-information, tailored by local experts, and leverages platforms frequently used by specific, we suggest that the benefits of such interventions are likely more durable than generic, one-size-fits-all approaches. While mis- and dis-information are pervasive and complex, individual choices regarding the consumption and sharing of information play a critical role in their spread. Carefully designed

interventions, such as the one examined in this study, can empower individuals to make more informed choices, mitigating the negative social consequences of misinformation.

References

Ajzen, I. (1991). The Theory of Planned Behavior. *Organizational Behavior and Human Decision Processes*, *50*, 179–211. https://doi.org/0749-5978/91

Akbar, S. Z., Panda, A., & Pal, J. (2022). Political hazard: Misinformation in the 2019 Indian general election campaign. *South Asian History and Culture*, *13*(3), 399–417. https://doi.org/10.1080/19472498.2022.2095596

Albee, J. J., Smith, M. L., Arnold, J. M., & Dennis, L. B. (2019). Digging Struggling Students Out of the Summer Reading Slump. *The Reading Teacher*, *73*(3), 291–299. https://doi.org/10.1002/trtr.1847

Altman, D. G., & Bland, J. M. (1994). Diagnostic tests. 1: Sensitivity and specificity. *BMJ : British Medical Journal*, *308*(6943), 1552.

Alwreikat, A. (2022). Sharing of Misinformation during COVID-19 Pandemic: Applying the Theory of Planned Behavior with the Integration of Perceived Severity. *Science & Technology Libraries*, *41*(2), 133–151. https://doi.org/10.1080/0194262X.2021.1960241

American Psychological Association. (2023, November). *Using psychological science to understand and fight health misinformation*. Https://Www.Apa.Org. https://www.apa.org/pubs/reports/health-misinformation

Anderson, J., & Rainie, L. (2017, October 19). The Future of Truth and Misinformation Online. *Pew Research Center*. https://www.pewresearch.org/internet/2017/10/19/the-future-of-truth-and-misinformation-online/

Anderson, K. (2019). Truth, Lies, and Likes: Why Human Nature Makes Online Misinformation a Serious Threat (and What We Can Do about It). *Law & Psychology Review*, *44*, 209.

Angeline, M., Safitri, Y., & Luthfia, A. (2020). Can the Damage be Undone? Analyzing Misinformation during COVID-19 Outbreak in Indonesia. *2020 International Conference on Information Management and Technology (ICIMTech)*, 360–364. https://doi.org/10.1109/ICIMTech50083.2020.9211124

Badrinathan, S., & Chauchard, S. (2024). Researching and countering misinformation in the Global South. *Current Opinion in Psychology*, *55*, 101733. https://doi.org/10.1016/j.copsyc.2023.101733

Bairoliya, N., & McKiernan, K. (2024). The welfare costs of misinformation. *Journal of Economic Dynamics and Control*, *169*, 104959. https://doi.org/10.1016/j.jedc.2024.104959

Basol, M., Roozenbeek, J., Berriche, M., Uenal, F., McClanahan, W. P., & Linden, S. van der. (2021). Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against COVID-19 misinformation. *Big Data & Society*, *8*(1), 20539517211013868. https://doi.org/10.1177/20539517211013868

Basol, M., Roozenbeek, J., & van der Linden, S. (2020). Good News about Bad News: Gamified Inoculation Boosts Confidence and Cognitive Immunity Against Fake News. *Journal of Cognition*, *3*(1), 2. https://doi.org/10.5334/joc.91

Batomen, B., & Benmarhnia, T. (2024). Staggered interventions with no control groups. *International Journal of Epidemiology*, *53*(6), dyae137. https://doi.org/10.1093/ije/dyae137

Baulch, E., Matamoros-Fernández, A., & Suwana, F. (2024). Memetic persuasion and WhatsAppification in Indonesia's 2019 presidential election. *New Media & Society*, *26*(5), 2473–2491. https://doi.org/10.1177/14614448221088274

BBC News. (2019, May 23). Indonesia post-election riots: Six dead in Jakarta as protesters clash with police. https://www.bbc.com/news/world-asia-48361782

Bernal, J. L., Cummins, S., & Gasparrini, A. (2017). Interrupted time series regression for the evaluation of public health interventions: A tutorial. *International Journal of Epidemiology*, *46*(1), 348–355. https://doi.org/10.1093/ije/dyw098

Braddock, K. (2022). Vaccinating Against Hate: Using Attitudinal Inoculation to Confer Resistance to Persuasion by Extremist Propaganda. *Terrorism and Political Violence*, *34*(2), 240–262. https://doi.org/10.1080/09546553.2019.1693370

Brehmer, M. (2023). Perceived Moral Norms in an Extended Theory of Planned Behavior in Predicting University Students' Bystander Intentions toward Relational Bullying. *European Journal of Investigation in Health, Psychology and Education*, *13*(7), Article 7. https://doi.org/10.3390/ejihpe13070089

Carthy, S. L., & Sarma, K. M. (2021). Countering Terrorist Narratives: Assessing the Efficacy and Mechanisms of Change in Counter-narrative Strategies. *Terrorism and Political Violence*, *0*(0), 1–25. https://doi.org/10.1080/09546553.2021.1962308

Chang, H.-C. H., Haider, S., & Ferrara, E. (2021). Digital Civic Participation and Misinformation during the 2020 Taiwanese Presidential Election. *Media and Communication*, *9*(1), 144–157. https://doi.org/10.17645/mac.v9i1.3405

Chen, J. J., & Krieger, N. J. (2023). Learning gain rather than learning loss during COVID-19: A proposal for reframing the narrative. *Contemporary Issues in Early Childhood*, *24*(1), 82–86. https://doi.org/10.1177/14639491211073144

Chen, L., & Fu, L. (2022). Let's fight the infodemic: The third-person effect process of misinformation during public health emergencies. *Internet Research*, *32*(4), 1357–1377. https://doi.org/10.1108/INTR-03-2021-0194

Cohen, J. (2013). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Compton, J., Ivanov, B., & Hester, E. (2022). Inoculation Theory and Affect. *International Journal of Communication*, *16*(0), Article 0.

Compton, J., Jackson, B., & Dimmock, J. A. (2016). Persuading Others to Avoid Persuasion: Inoculation Theory and Resistant Health Attitudes. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00122

Compton, J., van der Linden, S., Cook, J., & Basol, M. (2021). Inoculation theory in the post-truth era: Extant findings and new frontiers for contested science, misinformation, and conspiracy theories. *Social and Personality Psychology Compass*, *15*(6), e12602. https://doi.org/10.1111/spc3.12602

Cooke, R., Dahdah, M., Norman, P., & French, D. P. (2016). How well does the theory of planned behaviour predict alcohol consumption? A systematic review and meta-analysis. *Health Psychology Review*, *10*(2), 148–167. https://doi.org/10.1080/17437199.2014.947547

Facciani, M. J., Apriliawati, D., & Weninger, T. (2024). Playing Gali Fakta inoculates Indonesian participants against false information. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-152

Geldhof, G. J., Warner, D. A., Finders, J. K., Thogmartin, A. A., Clark, A., & Longway, K. A. (2018). Revisiting the utility of retrospective pre-post designs: The need for mixed-method pilot data. *Evaluation and Program Planning*, *70*, 83–89. https://doi.org/10.1016/j.evalprogplan.2018.05.002

Harjani, T., Basol, B., Melisa-Sinem, Roozenbeek, J., & van der Linden, S. (2023). Gamified Inoculation Against Misinformation in India: A Randomized Control Trial. *Journal of Trial & Error*, *3*(1). https://doi.org/10.36850/i3.1

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29–29. https://doi.org/10.1038/466029a

Howard, G. S., & Dailey, P. R. (1979). Response-shift bias: A source of contamination of self-report measures. *Journal of Applied Psychology*, *64*(2), 144–150. https://doi.org/10.1037/0021-9010.64.2.144

Huston, P., & Peterson, R. (2001). Withholding Proven Treatment in Clinical Research. *New England Journal of Medicine*, *345*(12), 912–914. https://doi.org/10.1056/NEJM200109203451210

IREX. (n.d.). *Learn to Discern: Media Literacy Trainer's Manual | IREX*. Retrieved February 17, 2025, from https://www.irex.org/resource/learn-discern-media-literacy-trainers-manual

Jeong, S.-H., Cho, H., & Hwang, Y. (2012). Media Literacy Interventions: A Meta-Analytic Review. *Journal of Communication*, *62*(3), 454–472. https://doi.org/10.1111/j.1460-2466.2012.01643.x

Kemp, S. (2023, February 9). *Digital 2023: Indonesia*. DataReportal – Global Digital Insights. https://datareportal.com/reports/digital-2023-indonesia

Kroke, A. M., & Ruthig, J. C. (2024). Conspiracy beliefs predicting health behaviors: An integration of the theory of planned behavior and health belief model. *Current Psychology*, *43*(9), 7959–7973. https://doi.org/10.1007/s12144-023-04953-y

Kruglanski, A. W. (with Bélanger, J. J., & Gunaratna, R.). (2019). *The three pillars of radicalization: Needs, narratives, and networks / Arie Kruglanski, Jocelyn J. Bélanger, Rohan Gunaratna.* Oxford University Press.

Kruglanski, A. W., Molinario, E., Ellenberg, M., & Di Cicco, G. (2022). Terrorism and conspiracy theories: A view from the 3N model of radicalization. *Current Opinion in Psychology*, *47*, 101396. https://doi.org/10.1016/j.copsyc.2022.101396

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, *28*, 1–26. https://doi.org/10.18637/jss.v028.i05

Lewandowsky, S., & van der Linden, S. (2021). Countering Misinformation and Fake News Through Inoculation and Prebunking. *European Review of Social Psychology*, *32*(2), 348–384. https://doi.org/10.1080/10463283.2021.1876983

Maertens, R., Roozenbeek, J., Basol, M., & van der Linden, S. (2021). Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied*, *27*(1), 1–16. https://doi.org/10.1037/xap0000315

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*, 61–83. https://doi.org/10.1016/j.jrp.2013.09.008

McDonnell, I., & MacKinnon, T. (2020). Case Study: Misinformation in Indonesia. *GeoPoll*. https://www.geopoll.com/misinformation-indonesia/

McGuire, W. J. (1961). Resistance to persuasion conferred by active and passive prior refutation of the same and alternative counterarguments. *The Journal of Abnormal and Social Psychology*, *63*(2), 326–332. https://doi.org/10.1037/h0048344

McGuire, W. J., & Papageorgis, D. (1961). The relative efficacy of various types of prior belief-defense in producing immunity against persuasion. *The Journal of Abnormal and Social Psychology*, *62*(2), 327–337. https://doi.org/10.1037/h0042026

Moore, D., & Tananis, C. A. (2009). Measuring Change in a Short-Term Educational Program Using a Retrospective Pretest Design. *American Journal of Evaluation*, *30*(2), 189–202. https://doi.org/10.1177/1098214009334506

Mujani, S., & Kuipers, N. (2020). Who Believed Misinformation during the 2019 Indonesian Election? *Asian Survey*, *60*(6), 1029–1043. https://doi.org/10.1525/as.2020.60.6.1029

Nurhayati-Wolff, H. (2023). *Indonesia: Share of male Instagram users by age 2024*. Statista. https://www.statista.com/statistics/997029/share-of-male-instagram-users-by-age-indonesia/

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867–872. https://doi.org/10.1016/j.jesp.2009.03.009

Piltch-Loeb, R., Su, M., Hughes, B., Testa, M., Goldberg, B., Braddock, K., Miller-Idriss, C., Maturo, V., & Savoia, E. (2022). Testing the Efficacy of Attitudinal Inoculation Videos to Enhance COVID-19 Vaccine Acceptance: Quasi-Experimental Intervention Trial. *JMIR Public Health and Surveillance*, *8*(6), e34615. https://doi.org/10.2196/34615

Polarization & Extremism Research & Innovation Lab. (2025). *Developing and Using Critical Comprehension (DUCC)*. PERIL Research. https://perilresearch.com/projects/ducc/

Prahassacitta, V., & Harkrisnowo, H. (2021). Criminal Disinformation in Relation to the Freedom of Expression in Indonesia: A Critical Study. *Comparative Law Review*, *27*(1), 135–165.

Pratt, C. C., McGuigan, W. M., & Katzev, A. R. (2000). Measuring Program Outcomes: Using Retrospective Pretest Methodology. *American Journal of Evaluation*, *21*(3), 341–349. https://doi.org/10.1177/109821400002100305

Pundir, V., Devi, E. B., & Nath, V. (2021). Arresting fake news sharing on social media: A theory of planned behavior approach. *Management Research Review*, *44*(8), 1108–1138. https://doi.org/10.1108/MRR-05-2020-0286

Quirk, S. (2021). Lawfare in the Disinformation Age: Chinese Interference in Taiwan's 2020 Elections. *Harvard International Law Journal*, *62*, 525.

Rahmawan, D., Garnesia, I., & Hartanto, R. (2024). Content analysis of MAFINDO's verified WhatsApp-related misinformation in Indonesia (July 2015–July 2020). Jurnal Kajian Jurnalisme, 8(1), 99–114. https://doi.org/10.24198/jkj.v8i1.54463

Rascoe, A. (2024, October 13). Misinformation and conspiracy theories about Hurricane Helene are spreading online. *NPR*. https://www.npr.org/2024/10/13/nx-s1-5148893/misinformation-and-conspiracy-theories-about-hurricane-helene-are-spreading-online

Roberts-Ingleson, E. M., & McCann, W. S. (2023). The Link between Misinformation and Radicalisation: Current Knowledge and Areas for Future Inquiry. *Perspectives on Terrorism*, *17*(1), 36–49.

Roozenbeek, J., van der Linden, S., Goldberg, B., Rathje, S., & Lewandowsky, S. (2022). Psychological inoculation improves resilience against misinformation on social media. *Science Advances*, *8*(34), eabo6254. https://doi.org/10.1126/sciadv.abo6254

Rosner, B., Glynn, R. J., & Lee, M.-L. T. (2006). The Wilcoxon Signed Rank Test for Paired Comparisons of Clustered Data. *Biometrics*, *62*(1), 185–192. https://doi.org/10.1111/j.1541-0420.2005.00389.x

Rulis, M. (2024). The Influences of Misinformation on Incidences of Politically Motivated Violence in Europe. *The International Journal of Press/Politics*, 19401612241257873. https://doi.org/10.1177/19401612241257873

Sailer, M., & Homner, L. (2020). The Gamification of Learning: A Meta-analysis. *Educational Psychology Review*, *32*(1), 77–112. https://doi.org/10.1007/s10648-019-09498-w

Schumann, S., & Barton, M. (2024). Does Attitudinal Inoculation Confer Resistance to Violent Extremist Propaganda? Assessing Mechanisms, Long-Term Effects, and the Advantage of Visuals. *Journal of Community & Applied Social Psychology*, *34*(6), e2898. https://doi.org/10.1002/casp.2898

Sharon, A. (2024, September 12). *Indonesia's Commitment to Inclusive Internet Access – OpenGov Asia*. https://opengovasia.com/2024/09/12/indonesias-commitment-to-inclusive-internet-access/

Syam, H. M., & Nurrahmi, F. (2020). "I Don't Know If It Is Fake or Real News" How Little Indonesian University Students Understand Social Media Literacy. *Jurnal*

Komunikasi: Malaysian Journal of Communication, *36*(2), Article 2. http://ejournal.ukm.my/mjc/article/view/36189

Theisen, W., Brogan, J., Thomas, P. B., Moreira, D., Phoa, P., Weninger, T., & Scheirer, W. (2021). Automatic Discovery of Political Meme Genres with Diverse Appearances. *Proceedings of the International AAAI Conference on Web and Social Media*, *15*, 714–726. https://doi.org/10.1609/icwsm.v15i1.18097

Traberg, C. S., Roozenbeek, J., & van der Linden, S. (2022). Psychological Inoculation against Misinformation: Current Evidence and Future Directions. *The ANNALS of the American Academy of Political and Social Science*, *700*(1), 136–151. https://doi.org/10.1177/00027162221087936

Tversky, A., & Kahneman, D. (1986). Rational Choice and the Framing of Decisions. *The Journal of Business*, *59*(4), S251–S278.

Van der Wal, Z. (2015). "All quiet on the non-Western front?" A review of public service motivation scholarship in non-Western contexts. *Asia Pacific Journal of Public Administration*, *37*(2), 69–86. https://doi.org/10.1080/23276665.2015.1041223

Wang, Y., McKee, M., Torbica, A., & Stuckler, D. (2019). Systematic Literature Review on the Spread of Health-related Misinformation on Social Media. *Social Science & Medicine*, *240*, 112552. https://doi.org/10.1016/j.socscimed.2019.112552

Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policy making (2017)*. https://policycommons.net/artifacts/421935/information-disorder/1392979/

World Economic Forum. (2024, January 16). *Special Address by Ursula von der Leyen, President of the European Commission*. World Economic Forum. https://www.weforum.org/meetings/world-economic-forum-annual-meeting-2024/sessions/special-address-by-ursula-von-der-leyen-president-of-the-european-commission-96293a5a9d/

Zhulfakar. (2024, September 12). *Indonesia's Internet Access Hits 79.5 Pct as Speed Rises Tenfold in a Decade*. Jakarta Globe. https://jakartaglobe.id/tech/indonesias-internet-access-hits-795-pct-as-speed-rises-tenfold-in-a-decade

About the JD Journal for Deradicalization

The JD Journal for Deradicalization is the world's only peer reviewed periodical for the theory and practice of deradicalization with a wide international audience. Named an "essential journal of our times" (Cheryl LaGuardia, Harvard University) the JD's editorial board of expert advisors includes some of the most renowned scholars in the field of deradicalization studies, such as Prof. Dr. John G. Horgan (Georgia State University); Prof. Dr. Tore Bjørgo (Norwegian Police University College); Prof. Dr. Mark Dechesne (Leiden University); Prof. Dr. Cynthia Miller-Idriss (American University Washington D.C.); Prof. Dr. Julie Chernov Hwang (Goucher College); Prof. Dr. Marco Lombardi, (Università Cattolica del Sacro Cuore Milano); Dr. Paul Jackson (University of Northampton); Professor Michael Freeden, (University of Nottingham); Professor Hamed El-Sa'id (Manchester Metropolitan University); Prof. Sadeq Rahimi (University of Saskatchewan, Harvard Medical School), Dr. Omar Ashour (University of Exeter), Prof. Neil Ferguson (Liverpool Hope University), Prof. Sarah Marsden (Lancaster University), Prof. Maura Conway (Dublin City University), Dr. Kurt Braddock (American University Washington D.C.), Dr. Michael J. Williams (The Science of P/CVE), Dr. Mary Beth Altier (New York University) and Dr. Aaron Y. Zelin (Washington Institute for Near East Policy), Prof. Dr. Adrian Cherney (University of Queensland), Dr. Wesley S. McCann (RTI International), and Dr. Daren Fisher (Hampton University).

For more information please see: www.journal-derad.com

Twitter: @JD_JournalDerad
Facebook: www.facebook.com/deradicalisation

The JD Journal for Deradicalization is a proud member of the Directory of Open Access Journals (DOAJ).

ISSN: 2363-9849

Editor in Chief: Daniel Koehler