Mandile, Simona

Working Paper

The Dark Side of Social Media: Recommender Algorithms and Mental Health

CESifo Working Paper, No. 11648

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Mandile, Simona (2025) : The Dark Side of Social Media: Recommender Algorithms and Mental Health, CESifo Working Paper, No. 11648, CESifo GmbH, Munich

This Version is available at:
https://hdl.handle.net/10419/314687

CESifo WORKING PAPERS

11648
2025

January 2025

The Dark Side of Social Media: Recommender Algorithms and Mental Health

Simona Mandile

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 11648*

The Dark Side of Social Media: Recommender Algorithms and Mental Health

Abstract

This paper investigates the impact of social media algorithms on mental health outcomes. I exploit a quasi-experimental setting combining data from the Dutch Longitudinal Internet Studies for the Social Sciences (LISS) coupled with the introduction of the algorithmic feed on Instagram in 2016. I estimate a differences-in-differences model comparing individuals having an Instagram account with individuals who have an account on social media platforms other than Instagram. Using a longitudinal dataset, allows for comparison of the same individuals before and after the introduction of the algorithm. The results show that the introduction of the algorithm on Instagram had a negative impact on teenagers mental health. Furthermore, I show that this effect cannot be attributed to a decrease in stigma surrounding mental health issues or an increased likelihood of individuals reporting such conditions. Additionally, evidence on mechanisms suggests that the results are due to the algorithm on Instagram favoring negative social comparisons.

JEL-Codes: I120, I310, L820, L860.

Keywords: social media, recommendation algorithm, mental health.

Simona Mandile
University of Bergamo / Italy
simona.mandile@unibg.it

January 20, 2025
I am indebted to Francesco Sobbrio for his guidance and support during my entire PhD path. This work benefited from valuable comments from Giorgio Gulino, Vincenzo Atella, Emilio Calvano, Elisa Facchetti, Paolo Pinotti, Maria Petrova, Libertad Gonz´alez and conference participants to the CESifo Area Conference on Economics of Digitization 2024. I am also thankful to participants of the CLEAN Workshop (Bocconi), Tor Vergata Reading Group in Political Economy & Labour and PhD Forum (Tor Vergata University of Rome) for their feedback. All remaining errors are my own.

1 Introduction

Mental health problems can be extremely damaging to individuals, families, and communities. They also place a significant burden on societies and economies, with the economic costs reaching as high as 4% of GDP (OECD (2022)). Additionally, people with mental illness tend to have worse educational, job, criminal and physical health outcomes compared to those with good mental health (Currie and Stabile (2006); Biasi et al. (2021); Anderson et al. (2015); Haushofer and Fehr (2014)). At the same time that social media started gaining popularity in the mid-2000s, the mental health of young people began to deteriorate (Patel et al. (2016)).[1] Although the ultimate causes of this phenomenon are still uncertain, this trend is often linked to the widespread use of the Internet and social media, which have significantly changed the way people spend their time and connect with each other (Twenge and Campbell (2019); Castellacci and Tveito (2018); Braghieri et al. (2022)). Concerns about the negative impact of digital technologies on mental health are further supported by industry insiders like Frances Haugen, a former Facebook employee. She leaked internal documents to the Wall Street Journal and the Securities and Exchange Commission that suggested Facebook knew that the use of Instagram may hurt the mental health of young women and girls.[2]

There is evidence consistent with the hypothesis that Internet and social media are partly responsible for the recent deterioration in mental health among teenagers and young adults (Donati et al. (2022); Braghieri et al. (2022)). However, well-identified causal evidence on which social media features play a major role in shaping the mental health of users of these platforms is scarce. Social media platforms have undergone dramatic transformations in recent years, particularly with the shift away from chronological user feeds. Initially, posts were displayed in the order they were published, but the advent of algorithmic recommender systems fundamentally altered this approach. Facebook is often credited with pioneering algorithmic recommendation on social media in 2011, a model later adopted by platforms such as Instagram, Twitter, and the more recent TikTok. Recommender algorithms were introduced to enhance user engagement and optimize

[1] In 2023, more than half of the world's population used social media, and the average person spent about two and a half hours each day on social media platforms (We Are Social 2024)

[2] Go to `https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739`

the overall platform experience. These algorithms personalize content delivery based on users' interests, behaviors, and interactions, making platforms more engaging by presenting content that aligns closely with individual preferences. This personalization not only increases user interaction with the platform but also extends the time users spend on it and encourages frequent returns. At the same time, social media platforms may foster digital addiction (Allcott et al. (2022)) and create negative externalities (Bursztyn et al. (2023)). Accordingly, there are public opinion concerns regarding the potential adverse consequences of social media recommender algorithms, including negative mental health effects, the spread of misinformation, and the reinforcement of echo chambers and filter bubbles. These algorithms prioritize content that aligns with users' existing opinions and preferences, potentially exacerbating these issues.[3]

In this paper, I provide quasi-experimental evidence of the impact of social media algorithms on mental health by exploiting the introduction of the algorithm on Instagram in 2016. Instagram was established in 2010 and has rapidly grown to 1.65 billion users at the start of 2024 (We Are Social, 2024). Before 2016, Instagram showed users' feeds with posts exclusively from accounts they followed, arranged in the order of when they were shared, with the most recent posts at the top. Starting from 2016, Instagram introduced an algorithmic system to curate content for its users and to give them "what they want to see".[4] Specifically, the feed still showcases content from users followed by individuals, but the order is now determined by the algorithm's assessment of the user's preferences rather than being purely chronological. Various factors are considered by the algorithm to determine the content shown to users, with one significant factor being the level of interaction a post receives. Posts with higher engagement, such as likes, comments, and shares, are more likely to be promoted and displayed to a broader audience. Additionally, post tags provide Instagram with insights into the target audience or individuals who may be interested in viewing the post (Agung and Darma (2019)).

I exploit the rich information provided by the the Dutch Longitudinal Internet Studies

[3]See among others:: https://www.wsj.com/story/tiktok-floods-teenagers-with-eating-disorder-videos-b20c2c73; https://integrityinstitute.org/blog/misinformation-amplification-tracking-dashboard; https://www.forbes.com/sites/traversmark/2023/12/09/how-social-media-uses-the-baader-meinhof-phenomenon-to-lie-convincingly/?sh=3462ebfa2f78; https://www.wired.com/story/meta-social-media-polarization/

[4]https://www.wired.com/2016/03/instagram-will-soon-show-thinks-want-see/

for the Social Sciences (LISS) panel, which provides information on a wide range of areas, including mental health and social media use. I employ a difference-in-differences empirical strategy and find that the introduction of the algorithm on Instagram had a negative impact on teenagers mental health. I employ a difference-in-differences empirical design comparing Instagram users after versus before the introduction of the algorithmic feed with respect to users of other social media after versus before such a change. The results show a robust negative impact of the introduction of Instagram recommender algorithm on the mental health of teenagers. Having a longitudinal dataset, I can compare the same individuals before and after the introduction of the algorithm. Comparing these two groups allows me to obtain causal estimates of the introduction of the algorithm on Instagram on individuals mental health. I provide evidence to support the underlying parallel trend assumption by presenting event study estimates showing the absence of pre-trends. This empirical strategy allows me to rule out differences across time that affect all individuals in a similar way, such as certain macroeconomic fluctuations and individual-specific differences fixed in time (e.g., poorer individuals may have worse baseline mental health than richer individuals). I find that the estimated poor mental health index for teenagers increased by 0.394 standard deviation units following the introduction of the algorithm on the social media platform. This magnitude is almost the same as the effect of losing one's job on mental health, as reported by Paul and Moser (2009). Moreover, the magnitude found in this work is around four times larger than that reported in Braghieri et al. (2022), which measures the effect of the introduction of Facebook at the college level on students' mental health. First, these differences may stem from the distinct content and design features of Instagram compared to the early version of Facebook. Second, the amplified impact can be attributed to significant technological advancements in social media platforms over the past 15 years. Additionally, the widespread adoption of smartphones, which enable constant connectivity regardless of time or location, offers another explanation for the observed differences in magnitude.

Furthermore, I present additional findings that offer deeper insight into the detrimental effects of social media algorithms on teenagers' mental health. First, the negative effects on mental health are particularly pronounced among first-generation immigrants and male teenagers who report poorer relationships with their parents. These characteris-

tics are well-documented predictors of increased susceptibility to mental health challenges, suggesting that the algorithm's impact may be especially harmful to already vulnerable groups. Second, the study reveals that teenagers experience significant downstream consequences stemming from their emotional distress. Following the introduction of Instagram's algorithm, these individuals are more likely to report that their emotional difficulties have directly interfered with their daily activities, social interactions, and work or study performance. Notably, following the introduction of the algorithm, male adolescents, in particular, have significantly reduced the time spent socializing with friends, family, neighbors, and in communal spaces such as cafes and bars. Similarly, girls have reported a decline, particularly in the quality of their relationships. This suggests that the algorithm not only adversely affects teenagers' mental well-being but also disrupts their daily lives and potentially undermines their long-term development.

The mechanism that seems to best explain the effect of social media and algorithms on mental health is negative social comparison. In fact, I find that the introduction of the algorithm on Instagram affects individuals' need for social validation, self-esteem and self-worth. Specifically, I find that after 2016, female teenagers' need for social validation increases and self-esteem and self-worth decreases, while there is no effect for male teenagers. I also show that the introduction of the algorithm on Instagram affected more severely the mental health of teenagers who might be more likely to be affected by unfavorable social comparisons. Individuals exhibiting traits such as envy of others' successes, discomfort with attention, and reluctance to engage socially are often more sensitive to how they perceive themselves in relation to others. These individuals may interpret others' achievements, attractiveness, or social connections as amplifying their own perceived shortcomings, making them especially vulnerable to the algorithm's effects. Finally, I present indirect evidence supporting the negative social comparison mechanism by showing that, following the introduction of the algorithm, teenagers increasingly perceive social media as harmful to their real-life social interactions. This finding aligns with recent empirical evidence highlighting the existence of a "social media trap" (Bursztyn et al. (2023)), i.e., the coexistence, in the case of social media, of a large individual consumer surplus and negative product welfare. The social media trap is that users would rather the platforms did not exist, but fail to coordinate to stop using them. Bursztyn et al. (2023) show that

the primary reason active users continue using social media, despite preferring a world without them, is the fear of missing out (FoMO). Importantly, evidence suggests that social comparison is associated with higher levels of FoMO (Burnell et al. (2019)). As for other channels, I find no significant evidence that the negative effects of Instagram's algorithm on mental health stem from increased or disruptive internet use. Specifically, there is no evidence that individuals are spending more time online across different devices. Furthermore, I observe no substantial rise in engagement with various online activities, except for an increase among female teenagers in the time spent downloading software, music, and films. In fact, I provide evidence suggesting a decline in time spent on certain activities. For instance, female teenagers appear to be spending less time on newsgroups, forums, blogs and dating, while male teenagers are showing a reduced engagement with blogs and newspapers. Additionally, there is no evidence consistent with a mechanism of reduced social stigma which might have increased individuals' willingness to report experiencing mental health issues. Finally, I show that the observed effects are not driven by changes in other behaviors known to impact mental health, such as substance use, alcohol consumption and practicing physical activity.

My paper contributes to the growing economic literature on the impact of the Internet and social media on mental health by focusing on one of the most revolutionary features of social media platforms, i.e., recommender algorithms. I mainly refer to the existing literature on the effects of the internet and specifically of social media on mental health (Donati et al. (2022); Arenas-Arroyo et al. (2022); Golin (2022); Braghieri et al. (2022); Allcott et al. (2020); Allcott et al. (2022); Mosquera et al. (2020)). Braghieri et al. (2022) is the paper which is closest to mine, as it exploits a quasi-experimental setting to provide causal evidence on both direct and indirect effects of social media on mental health. They focus on the entire population of students taking the National College Health Assessment (NCHA) survey [5], including both students who did and who did not have a Facebook account. Therefore, they cannot disentangle the direct effect of having a Facebook account from the indirect effect on students who did not join the platform, but whose peers did. Differently from Braghieri et al. (2022), I have detailed information on individual social

[5]The National College Health Assessment (NCHA) survey is the most comprehensive survey about student mental and physical health available at the time of Facebook's expansion.

media usage so I can measure the direct effect of having an Instagram account. Moreover, while their analysis examines the pre-recommender algorithm era of social media by focusing on the effect of Facebook's introduction on mental health, my study complements theirs by investigating the impact of the introduction of Instagram's recommender algorithm on mental health. Finally, their analysis examines up to two years following the introduction of Facebook, while my data extends to five years after the algorithm's implementation, allowing me to explore long-term effects. Allcott et al. (2020), Allcott et al. (2022) and Mosquera et al. (2020) are experimental works that incentivize participants to reduce their social media use. They find negative effects of social media use on well-being, whereas Allcott et al. (2022) provides evidence of digital addiction. Differently from this experimental literature, I have a longitudinal panel dataset so I can measure long term effects. Moreover, by exploiting a quasi-experimental setting my estimates are less affected by experimenter demand, Hawthorne, and income effects. Indeed, papers listed above focus on treated individuals receiving compensation for reducing their social media usage, which means they are aware of their participation status. This awareness could potentially influence their behavior due to experimenter demand effects. Additionally, by being observed, they might change their own behaviors, leading to general Hawthorne effects. Furthermore, the incentive payments may directly influence self-reported well-being. Another issue with social media experiments is the use of selective samples, as they often screen out participants who do not meet specific criteria. Moreover, differently from these papers, rather than studying the partial equilibrium effects of paying individuals to reduce social media use, my estimates capture the general equilibrium effects of introducing algorithms to social media. Such general equilibrium effects are likely to be particularly important for technologies such as social media that have strong network externalities (Bursztyn et al. (2023)).

Secondly, I contribute to the literature on the specific effects of social media algorithms on various outcomes (Guess et al. (2023); Nyhan et al. (2023); Levy (2021); Huszár et al. (2022); Germano et al. (2022)). For example, Germano et al. (2022) show that social media algorithms designed to prioritize engagement foster misinformation and polarization, underscoring the unintended consequences of platform design on public discourse. Similarly, Huszár et al. (2022) exploits a similar quasi-experimental setting to mine, the introduc-

tion of the algorithm on Twitter, to evaluate its effects on political content. They find that the political right enjoys higher amplification compared to the political left. Guess et al. (2023) show that moving users out of algorithmic feeds influenced users' experiences on social media but it did not significantly alter levels of issue polarization, affective polarization, political knowledge, or other key attitudes. My contribution builds on this literature by investigating the effect of Instagram's algorithm, introduced on one of the most popular social media platforms among young people, on mental health outcomes, using a quasi-experimental framework.

Additionally, I also contribute to the existing literature on the determinants and consequences of mental illness (Ridley et al. (2020); Paul and Moser (2009); Haushofer and Shapiro (2016); Persson and Rossin-Slater (2018); Golberstein et al. (2019)). I add to this body of research by investigating the impact of social media, with a particular focus on their algorithms, which many believe play a significant role in the recent increase in depression rates among teenagers and young adults (Twenge and Campbell (2019)).

Finally, I provide suggestive evidence on the external validity of the results by exploiting repeated cross-sectional data from the Pew Research Center in the US employing a similar difference-in-differences design. Specifically, I show that, in the US context, the introduction of Instagram's recommender algorithm was associated with a higher likelihood of individuals reporting that social media made them feel worse about their own lives. It also increased the pressure to post content that portrays them positively and heightened the importance placed on receiving likes and comments.

The rest of the article is organized as follows. Section 2 provides the background. Section 3 presents the Data. Section 4 presents the empirical strategy employed. Section 5 reports the results. Section 6 investigates the potential mechanisms underlying the results. Section 7 provides supporting evidence that the identified effects are not specific to the Netherlands but likely to apply more widely. Section 8 concludes.

2 Background

2.1 Mental Health

Mental illnesses, such as depression, anxiety, bipolar disorder, and schizophrenia, are very common conditions. Data from Eurofound's e-survey indicate that in the spring 2022, the 55% of the EU population could be considered at risk of depression. The percentage of people at risk of depression ranges from about 40 percent in Slovenia, Denmark and the Netherlands to about 65 percent in Poland, Greece and Cyprus. Mental illness affects people's lives, limiting their ability to study, work and be productive. The OECD estimates that the economic cost of mental health, due to treatment costs but also reduced productivity and lower employment, is about 4 percent of worldwide GDP (OECD (2022)).

According to the National Health Survey conducted in 2021, the prevalence of mental health issues among young individuals aged 12 to 24 in the Netherlands is 18 percent. Figure A1 shows a rise in the percentage of individuals 16 years of age or older who score less than 60 on the Mental Health Inventory (MHI).[6] Notably, the age groups experiencing the most significant increase in mental health issues are teenagers aged 16 to 20 and young adults aged 20 to 30. For those aged 16 to 20 in particular, the decline in mental health outcomes appears to begin around 2016, the same year Instagram introduced its algorithm.

2.2 Instagram and its Algorithm

Instagram is a free, online photo-sharing application and social network platform. It was founded in 2010 and acquired by Facebook in 2012. Instagram has experienced remarkable growth, reaching a user base of 1.65 billion at the start of 2024 (We Are Social, 2024). The primary purpose of this social media platform is to facilitate the sharing of photographs and short videos among its users.

Instagram has two main sections: the feed and the explore page. Before 2016 the user's feed displayed posts shared only by their *followings* (i.e., people and pages that they follow) in reverse chronological order, meaning the most recent posts appeared at the top. However, starting in March 2016, Instagram implemented an algorithmic approach

[6]The Figures refer to the 'Mental Health Inventory 5' or 'MHI-5'. It is an international standard for a specific measurement of mental health, consisting of 5 questions. For a detailed description of the outcome, treatment, and control variables, see Table A16.

to curate content for its users. The algorithm aimed to present users with the content they would be most interested in seeing. As a consequence, starting from 2016, the feed would still show users content posted by their followings, but the order was no longer purely chronological but determined by the algorithm's assessment of user preferences. The algorithm takes into account various factors to determine the content shown to users. One crucial factor is the level of interaction a post receives. The more interactions, such as likes, comments, and shares, a post has, the more likely it is to be promoted and displayed to a wider audience. Additionally, the tags of a post provide Instagram with information on the target audience or individuals who may be interested in viewing the post. This marked the beginning of a broader transformation for Instagram. Alongside the algorithm introduced in March 2016, Instagram launched the Stories feature in August of the same year, allowing users to share photos and videos that lasted only 24 hours. The introduction of Stories was part of a broader strategy aimed at enhancing user engagement and diversifying content formats. The placement and visibility of Stories prioritize content from accounts users interact with most frequently, further reinforcing Instagram's commitment to a personalized and tailored user experience.

3 Data

3.1 Dutch Longitudinal Internet Studies for the Social Sciences (LISS) panel

My analysis relies on the Dutch Longitudinal Internet Studies for the Social Sciences (LISS) panel (Scherpenzeel (2018)). It is a comprehensive research project that involves 5,000 households, encompassing approximately 7,500 individuals. This panel is based on a true probability sample of households, selected from the population register by Statistics Netherlands. It is a longitudinal study that is repeated on an annual basis. I specifically focus on the time period that spans from 2012 to 2021. The survey inquires about demographics, social media usage patterns, mental health, internet usage habits, personality traits, physical health status, alcohol and drug consumption. Remarkably, the LISS panel allows me to make dual contributions to the existing literature. Firstly, differently from the existing experimental works that mainly focus on the short-term, the LISS survey is a

panel dataset that allows me to measure long term effects. Secondly, differently from the existing quasi-experimental literature, the availability of detailed information on individual social media usage within the LISS panel dataset allows for the measurement of the direct effect of having an Instagram account.

To provide structure to my analysis and address concerns about multiple hypothesis testing, I organize individual mental health variables into nested groups and combine them into indices. The construction of the indices follows the methodology detailed in Braghieri et al. (2022). Firstly, I combined all mental health questions to form an index of overall poor mental health. The second level of analysis separates symptoms of mental illness (index of symptoms of poor mental health) from self-reported utilization of depression-related services (index of depression services) into distinct families. The third level of analysis further divides symptoms of mental illness into depression-related symptoms (index of depression symptoms) and symptoms associated with other mental health conditions (index of symptoms of other mental health conditions). Finally, I also considered individual variables themselves. The index of depression symptoms includes questions that inquire about various symptoms of depression, such as feeling anxious, very sad, depressed and gloomy. The index of symptoms of other mental health conditions actually coincide with an index of eating disorder related issues because it considers whether an individual suffers from anorexia. The overall index of symptoms of poor mental health encompasses both sets of symptoms. The index of depression services comprises questions inquiring whether an individual takes medicine for anxiety or depression, sleeping problems or whether the individual was in therapy for depression in the year of the survey. My indices are constructed as follows: first, I align all variables within an index so that higher values consistently indicate worse mental health outcomes. Second, I standardize these variables using means and standard deviations from the preperiod. Third, I calculate an equally weighted average of the index components, excluding observations with missing components from the analysis. Fourth, I standardize the final index. Consequently, my indices represent z-scores.

3.2 Descriptive Statistics

Table A1 presents descriptive statistics for all individuals included in the survey. Panel A of Table A1 shows that individuals with an Instagram account are younger, less wealthy and are less educated. Moreover, within those having an Instagram account there are more female and non-dutch individuals. Panel B of Table A1 shows that individuals with an Instagram account have worse baseline mental health outcomes than individuals without an Instagram account but they are less likely to take up depression services. Table A2 presents descriptive statistics only for my main group of interest, teenagers, namely those that belong to the cohort of individuals who were born between 1995 and 2000. Panel A of Table A2 shows that teenagers with an Instagram account are more likely to be female and with an higher net income. Panel B of Table A2 shows that teenagers with an Instagram account have worse baseline mental health outcomes than individuals without an Instagram account. The baseline differences across these two groups may lead one to wonder about the plausibility of the parallel trends assumption in this setting; I address concerns related to parallel trends in Section 5.

4 Empirical Strategy

I can exploit the introduction of the Instagram's algorithm in 2016 as a quasi-experimental variation to determine the causal effect of social media algorithms on mental health outcomes. To do this, I employ a difference-in-differences approach, which involves comparing the changes in outcomes before and after the introduction of the algorithm on Instagram. Specifically, I compare individuals who exclusively use Instagram, as well as those who use Instagram alongside other social media platforms, to a control group of individuals engaged in other social media platforms but not in Instagram (i.e., Facebook, Twitter and Youtube).

Firsly, I estimate the following model:

$$Y_{it} = \alpha_i + \delta_t + \beta \cdot Instagram_i \cdot Post_t + X_{it} \cdot \gamma + \epsilon_{it} \tag{1}$$

Where Y_{it} represents an outcome for individual i at time t. β is the coefficient of inter-

est since it identifies the average treatment effect on the treated (ATT) of the introduction of the Instagram's algorithm on individual mental health. $Instagram_i$ indicates if an individual has an Instagram account and $Post_t$ is an indicator for the post-treatment period. X_{it} corresponds to individual time-varying controls and standard errors are clustered at the individual level. Finally, α_i and δ_t are individual and time fixed effects. In this way, I can rule out that the results are driven by mental health outcomes evolving over time in a way that is common across individuals and by individual-specific differences fixed in time. The construction of my treatment indicator is straightforward. An individual is considered treated if, by 2016, they have an Instagram account; individuals who create an Instagram account after 2016 are excluded from the sample.[7] Treated individuals include those who solely use Instagram as well as those who have Instagram accounts alongside accounts on other social media platforms. The control group consists of individuals who, by 2016, have accounts on other social media platforms (i.e., Facebook, Twitter, YouTube) but do not have an Instagram account. I exclude from the sample individuals who create their social media accounts only after 2016 to maintain consistency with the construction of the treated group. By comparing these two groups of individuals who are active on at least one social media platform prior to Instagram's algorithm introduction, I can be more confident that the assumption of parallel trends holds. Before the algorithm was introduced, both the control and treatment groups should have exhibited similar trends in mental health. To ensure the plausibility of the parallel trend assumption, I estimate a dynamic version of Equation 1 and examine any potential pre-existing trends.

5 Results

5.1 Baseline Results

Table 1 shows results from equation 1 on the general index of poor mental health. I found that the introduction of the algorithm on Instagram had a negative impact on

[7] In the baseline, I exclude from the sample all individuals who create a social media account after 2016. However, when using the robust estimator proposed by De Chaisemartin and d'Haultfoeuille (2024), I include those individuals who create an Instagram account after 2016, as well as those who choose to leave Instagram, addressing switches in treatment status over time. This estimator effectively accounts for such dynamics and offers a robust framework for analyzing treatment effects in the presence of time-varying treatment assignments (Figure A3).

teenagers mental health, while not affecting other cohorts. Column 1 shows results for the specification in which I included individual and time fixed effects. In column 2, I also include controls which consists of net income, level of education, housing situation and level of urbanization. I find that the effects of the introduction of the algorithm on Instagram on mental health outcomes is statistically significant only for individuals born between 1995 and 2000. The effect size for these individuals in my preferred specification, namely the one that includes individual time-varying controls, is 0.394 standard deviation units. In the remainder of the work I will use this specification and focus only on teenagers.

Table 1: Effect of the Instagram's algorithm on the Index of Poor Mental Health

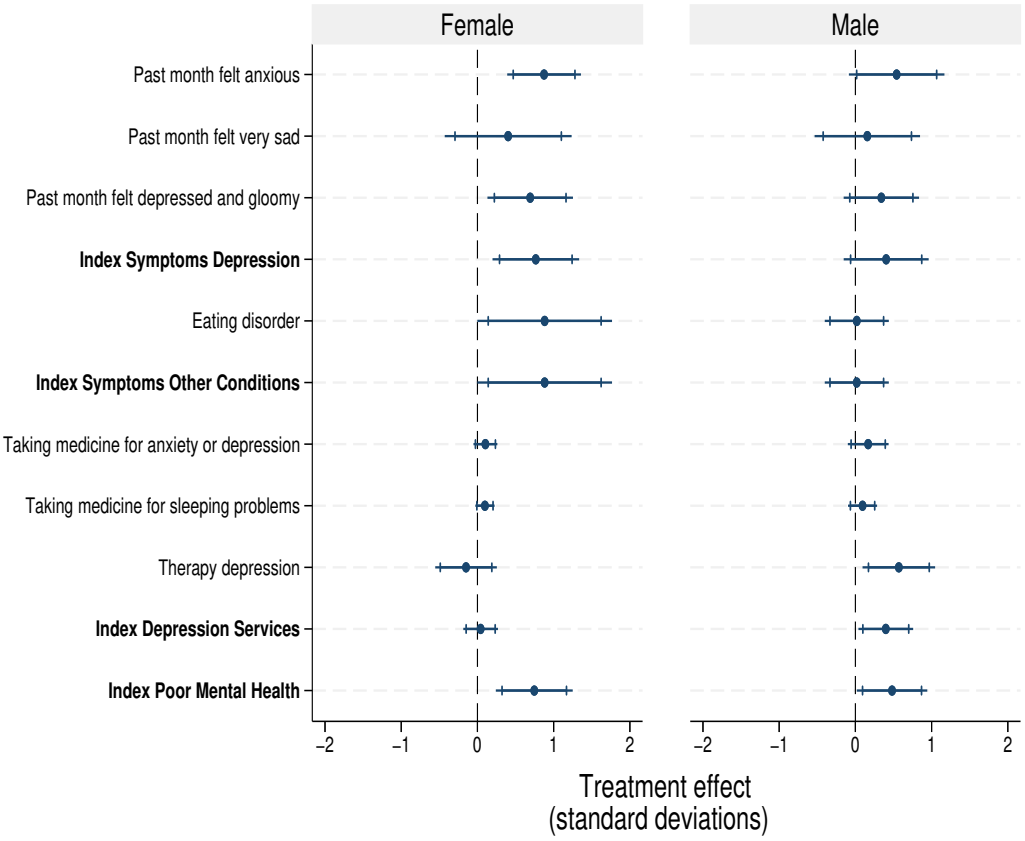
	Index of Poor Mental Health	
	(1)	(2)
Cohort 1995-2000:		
Post IG Algorithm Introduction	0.277**	0.394**
	(0.140)	(0.181)
Cohort 1981-1994:		
Post IG Algorithm Introduction	0.038	0.059
	(0.059)	(0.065)
Cohort 1965-1980		
Post IG Algorithm Introduction	-0.039	-0.048
	(0.055)	(0.056)
Individual fixed effects	✓	✓
Time fixed effects	✓	✓
Controls		✓
Observations	8,871	8,211

Notes. This table explores the effect of the introduction of the algorithm on Instagram on individuals' mental health. It presents estimates of coefficient β from equation 1 with my index of poor mental health as outcome variable. The index is standardized so that, in the preperiod, it has a mean of zero and a standard deviation of one. Column 1 estimates equation 1 without including controls column 2 estimates equation 1 including controls. My controls consist of net income, level of education, housing situation and level of urbanization. For a detailed description of the outcome, treatment, and control variables, see Table A16. Standard errors in parentheses are clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Comparing my findings to those in related work by Paul and Moser (2009), I observe that the impact of Instagram's algorithm introduction on teenagers is nearly equivalent to the effect of job loss. Moreover, it is over four times greater than the impact reported by Braghieri et al. (2022), which examines how the introduction of Facebook at the university level influenced students' mental health. The difference in the impact observed in my work compared to that reported in Braghieri et al. (2022) may arise from several key factors. First, it could stem from the distinct content and design features of Instagram

compared to the early version of Facebook. Second, the amplified impact can be linked to the substantial technological advancements in social media platforms over the past 15 years. Moreover, the widespread adoption of smartphones, enabling constant connectivity regardless of time or location, provides another possible explanation for the observed differences in magnitude.

Figure 1: Effects of the Instagram's algorithm on teenagers mental health by gender



Notes. This Figure explores the effects of the introduction of the Instagram's algorithm on all my mental health outcome variables and on the related indices, by gender. It displays estimates of coefficient β from equation 1 using my preferred specification, namely the one including controls and individual and time fixed effects. The outcome variables are my overall index of poor mental health, the individual components of the index, and three subindices: the index of depression symptoms, the index of symptoms of other mental health conditions, and the index of depression services. All outcomes are standardized so that, in the preperiod, they have a mean of zero and a standard deviation of one. My controls consist of net income, level of education, housing situation and level of urbanization. For a detailed description of the outcome, treatment, and control variables, see Table A16. The bars represent 95 and 90 percent confidence intervals. Standard errors are clustered at the individual level.

Figure 1 shows results on all individual outcome variables for female and male teenagers separately. For almost all outcomes, the point estimates are positive, indicating a decline in mental health. The effect is more pronounced for female teenagers and is driven by increased levels of anxiety, depression and gloominess and eating disorder. Although the

effect is smaller for male teenagers, they too show negative impacts from the introduction of the algorithm, with higher levels of anxiety along with an increased uptake of therapy for depression. Figure A2 shows the same results but for the entire sample of teenagers.

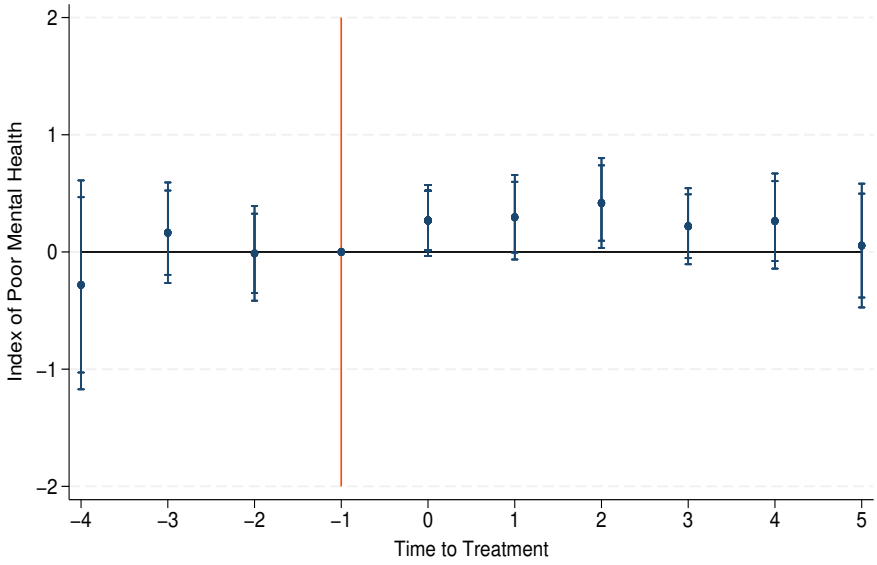
5.2 Event Study

To assess the plausibility of the parallel trends assumption, I estimate the following specification:

$$Y_{it} = \alpha_i + \delta_t + \beta_k \cdot \sum_{k=-4}^{5} \texttt{Instagram}_i + \epsilon_{it} \tag{2}$$

Where Y_{it} represents an outcome for individual i at time t and $Instagram_i$ indicates whether an individual has an Instagram account or not.

Figure 2: Event study for the Index of Poor Mental Health



Notes. This Figure overlays the event-study plot using a dynamic version of the TWFE model, equation 2. The outcome variable is my index of poor mental health. The index is standardized so that, in the pre-period, it has a mean of zero and a standard deviation of one. For a detailed description of the outcome, treatment, and control variables, see Table A16. The bars represent 95 and 90 percent confidence intervals. Standard errors are clustered at the individual level.

Figure 2 shows that the estimates are consistent with the parallel trend assumption. The coefficients before 2016, the year in which the algorithm was introduced, do not exhibit discernible pre-trends. The treatment effect increases in the first few years after the introduction of the algorithm, becoming significant in 2016, and remaining significant

until 2018. However, after 2019, the coefficient decreases and becomes non statistically significant. This may be due to different potential factors. First, other platforms might have pushed recommender algorithms operating in a similar direction. Second, individuals might have developed digital resilience[8], which enables individuals to adapt, manage, and thrive in the digital world while maintaining both security and well-being. Digital resilience combines digital wellbeing, such as fostering a healthy relationship with technology, limiting screen time, prioritizing mental and emotional health, and improving digital literacy, with effective security practices. Over time, users may become more familiar with the platform and learn strategies to minimize its negative effects on mental health. With years of experience, they can develop better tools and habits to navigate the platform more effectively, mitigating harmful effects that may be more intense for newer users. Third, the COVID-19 pandemic likely had a broad negative impact on mental health, affecting both the treatment and control groups. This shared adversity may have led to a "catch-up" effect, narrowing the gap between the two groups and contributing to a diminished treatment effect in 2020 and 2021.

I replicate my results using the robust estimator introduced by De Chaisemartin and d'Haultfoeuille (2024). In my baseline estimation, I focus exclusively on individuals who already had an Instagram account at the time the algorithm was introduced, excluding those who created accounts after 2016. This exclusion applies not only to treated individuals but also to those in the control group who later created an Instagram account. The robust estimator from De Chaisemartin and d'Haultfoeuille (2024) provides consistent estimates even when treatment effects vary over time. I extend my baseline analysis to include individuals who joined Instagram after 2016, defining the treatment as their first exposure to social media algorithms. This approach enables me to account for individuals who switch into the treatment after 2016, those who switch out, and those who never switch. Figure A3 shows qualitatively similar results to those obtained with the TWFE estimator.

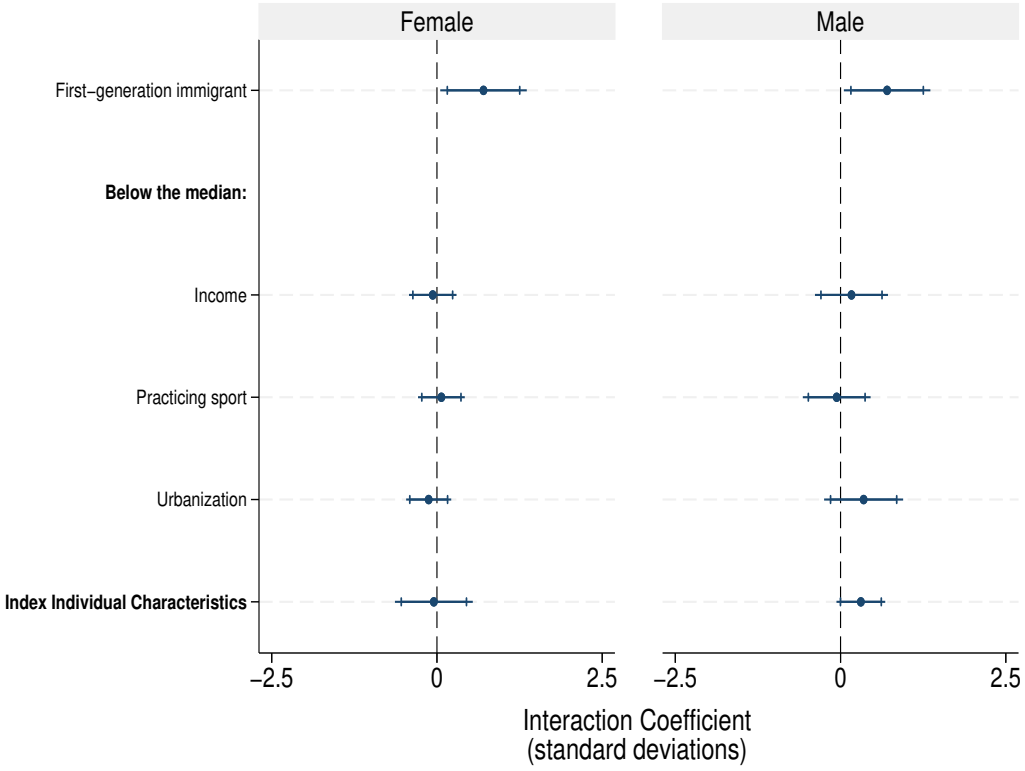
[8]Go to `https://digiwell.sk/wp-content/uploads/2024/01/DigiWELL-MM_final_ANG.pdf`

5.3 Heterogeneity

Figure 3 presents estimates of heterogeneous effects across various individual characteristics. The findings indicate that the introduction of the algorithm on Instagram has a more pronounced effect on first-generation immigrants. This aligns with evidence suggesting that first-generation immigrants are more prone to negative social comparisons than native-born individuals or second-generation immigrants. Specifically, first-generation immigrants may experience higher levels of social comparison due to factors like acculturation stress, limited social networks, and barriers to economic or social mobility (Tineo et al. (2024)). Adapting to a new culture can lead to feelings of inadequacy or insecurity, encouraging comparisons with natives or more integrated peers. Additionally, I find no significant differences when examining other dimensions of heterogeneity, such as income level, participation in sports, or the level of urbanization in their place of residence.

Figure A4 illustrates heterogeneous effects across various family characteristics. The findings reveal that, for male teenagers, the impact of the algorithm's introduction on Instagram is amplified when they have strained relationships with both their mother and father. A poor relationship with family members can intensify the adverse effects of external stressors, including social media algorithms, on mental health, as family typically serves as a key source of emotional support and stability. Teenagers growing up in dysfunctional family settings face an elevated risk of mental health issues, which, if untreated, can lead to lasting challenges such as depression and anxiety (Mphaphuli (2023)). Poor relationship quality within families is a significant stressor that can undermine well-being (Thomas et al. (2017)). Particularly for teenagers, a strong parental bond can provide a protective buffer against external pressures by fostering feelings of security, self-worth, and resilience. Without this foundational support, teenagers may seek validation from social media, which can heighten risks of social comparison, loneliness, and anxiety (Sela et al. (2020)). Studies show that teenagers in difficult family environments are more likely to rely on online interactions for acceptance and self-esteem, exposing them to the pitfalls of algorithm-driven content and increasing their vulnerability to negative social comparisons (Sela et al. (2020)).

Figure 3: Heterogeneous effects - individual characteristics



Notes. This Figure explores whether the effects of the introduction of the algorithm on Instagram on mental health are heterogeneous across a host of individual characteristics. Specifically, it presents estimates from a version of equation 1 in which my treatment indicator is interacted with various moderators. The outcome variable is my index of poor mental health. The index is standardized so that, in the pre-period, it has a mean of zero and a standard deviation of one. The moderators are indicators for: being a first-generation immigrant and being below the median income, engagement in sport activities and urbanization. All estimates are obtained using my preferred specification, namely the one including controls and individual and time fixed effects. My controls consist of net income, level of education, housing situation and level of urbanization. For a detailed description of the outcome, treatment, and control variables, see Table A16. The bars represent 95 and 90 percent confidence intervals. Standard errors are clustered at the individual level.

5.4 Downstream Effects

In this section, I examine whether the negative effects on mental health and emotional well-being caused by Instagram's algorithm interfere with individuals' ability to perform daily activities. The LISS survey includes questions that assess whether emotional difficulties and physical health issues impede activities such as (1) daily self-care tasks (e.g., going for a walk, climbing stairs, dressing, personal hygiene, and using the toilet); (2) social interactions (e.g., visiting friends and acquaintances); and (3) work or study-related tasks (e.g., job performance, housekeeping, or schoolwork).

To investigate the downstream implications of mental health, I estimate the impact on these outcomes by running separate analyses for individuals above and below the median

of my index measuring poor physical health. This approach allows me to identify whether the effect of emotional difficulties on these activities differs depending on physical health status. Indeed, for individuals below the median of poor physical health (i.e., those with relatively better physical health), any observed disruption in activities is more likely to be attributable to emotional difficulties rather than physical limitations. By focusing on this group, I can better capture the unique effect of emotional problems on daily functioning without the added complexity of physical health impairments that could also hinder these activities.

Figure A5 shows that the negative impact on mental health and emotional well-being from Instagram's algorithm extends to daily activities, social activities, and work or study-related activities for teenagers, reflected in my index of downstream effects. Notably, this effect is statistically significant only for individuals below the median of the poor physical health index, allowing me to "isolate" the effect of emotional difficulties specifically among those with relatively better physical health. These findings highlight that Instagram's algorithm exacerbates emotional problems to such an extent that they interfere with essential daily functioning and social engagement.

Building on these findings, the introduction of the algorithm appears to have broader consequences on offline interactions, as explored in Figures A6 and A7. Specifically, the disruptions to daily and social activities shown in Figure A5 are mirrored in patterns of diminished socialization. In Figure A6, I examine whether the algorithm influenced individuals' socialization patterns. One significant concern is that the algorithm may lead individuals to spend less time with family and friends, which could have profound implications, as insufficient social connection is strongly linked to long-term negative effects on both physical and mental health (Hawkley and Cacioppo (2010)). The figure demonstrates that the introduction of the algorithm is associated with a significant reduction in the amount of time teenagers, especially males, spend with family members, neighbors, friends outside their neighborhood, or visiting bars and cafes. This reduction in social time ties back to the downstream effects shown in Figure A5, reinforcing how the algorithm's impact cascades through both emotional well-being and social dynamics.

Figure A7 further illustrates the impact of the algorithm on various measures of social connection quality, such as satisfaction with social contacts, enjoyment of friendships,

feelings of connection, and experiences of loneliness or desertion. The results reveal a clear negative effect on these measures, particularly among female teenagers, where the effect is both negative and statistically significant for the index encompassing these variables. This indicates that the algorithm disrupts individuals' sense of social support and connection. This aligns with the evidence of downstream effects discussed earlier: the algorithm not only intensifies emotional problems but also weakens offline social interactions by diminishing the quality and quantity of social connections. Interestingly, Table A3 reveals that when examining activities less closely tied to social interactions, such as time spent doing sports, there are no significant effects. This lack of impact highlights that the observed negative consequences of Instagram's algorithm are predominantly confined to activities reliant on interpersonal connections. These findings strengthen the conclusion that the algorithm's introduction has downstream implications primarily affecting teenagers' offline interactions, rather than activities less dependent on social engagement.

5.5 Robustness Checks

I run some exercises to probe the robustness of my estimates. In 2016, Twitter implemented machine learning algorithms to curate tweets on the Home timeline using a personalized relevance model. This change meant users would see older tweets deemed relevant to them, along with some tweets from accounts they did not directly follow. Personalized ranking prioritizes certain tweets over others based on content features, social connections, and user activity. Importantly, unlike Instagram, Twitter used to provide users with the option to turn off the algorithm-based feed and revert to a chronological feed. Initially, this option was available in the account settings, allowing users to choose between "Show the best Tweets first" and a chronological feed. Many users expressed their disappointment on Twitter itself, using hashtags such as #RIPTwitter to complain about the change. Additionally, online petitions were created asking Twitter to maintain or restore the chronological feed. In response to the criticism, in 2018, Twitter made it even easier for users to switch between the algorithmic and chronological feeds by introducing an option directly in the feed. Users could click on a star in the top right corner of the feed to choose between "Home" (algorithmic) and "Latest Tweets" (chronological). In Table A4, I present the results of estimating equation 1, excluding individuals with a

Twitter account. Reassuringly, the results remain qualitatively similar to those reported in Table 1.

Secondly, in Table A5, I perform a placebo test to verify whether the observed effects are specific to the introduction of Instagram's algorithm or instead associated with the use of other social media platforms. Indeed, I replicate my baseline analysis, using different treatment definitions. Specifically, I present estimates of coefficient β from equation 1, using different groups of treated and control units. Specifically, I define the treatment as having a Facebook account compared to having a social media account other than Facebook. Another estimation defines the treatment as having a Twitter/X account compared to having a social media account other than Twitter/X. A third estimation defines the treatment as having a YouTube account compared to having a social media account other than YouTube. In all three exercises I consistently exclude individuals with Instagram accounts. The intuition behind these exercises is that if my results are driven by the introduction of the algorithm on Instagram and its effect on its users, I should not see any significant difference between the groups considered in this specification. Indeed, coefficients reported in Table A5 are all non significant.

As an additional test, Table A6 presents a placebo checks on an index of all physical rather than mental health outcomes in my dataset. Consistent with intuition, I find no statistically significant effect of the introduction of the algorithm on physical health outcomes.

Furthermore, I show that my results are not driven by the way the index is constructed. Table A7 exhibits results from equation 1 in which the dependent variable is an index of poor mental health that excludes observations for which some of the component variables are missing and an inverse-covariance weighted index that assigns a smaller weight to strongly correlated components (Anderson (2008)). Table A7 shows that the results remain qualitatively similar using the alternative index.

Next, Figure A8 shows that the results are not driven by any outcome variable. Indeed, I exploit different versions of the main index of poor mental health, each time excluding a different individual variable. I show that my estimates are robust to separately dropping each individual component of the index of poor mental health.

There could be concerns that Instagram users may inherently differ from other social

media users in ways that affect mental health. In Table A8 I address the presence of potential self-selection bias, by selecting a set of variables that account for both factors of vulnerability to mental health issues and traits that may incline individuals to use Instagram. Variables such as financial expectations for the future, satisfaction with appearance, and sense of equal worth to others capture aspects of emotional vulnerability and self-esteem, both of which can influence how individuals engage with social media content, particularly on platforms like Instagram. Moreover, the general index of poor mental health serves as a baseline measure that can signal any pre-existing mental health difficulties, allowing for a clearer understanding of mental health trajectories in relation to Instagram usage. Additionally, variables such as perceived job prospects in the coming year, self-perception as the "life of the party", and comfort around others provide insight into extroversion and sociability, traits that may make people more likely to engage with Instagram as a medium for social connection. Conversely, traits like tendency to be reserved or inclination to remain in the background offer a counterbalance, capturing introversion-related characteristics that can influence both the likelihood of Instagram use and psychological responses to social media engagement. I conduct a balance test on these characteristics in the pre-period between Instagram users and other social media users to assess potential self-selection bias. This approach allows me to determine whether there are any pre-existing differences that could simultaneously influence both the likelihood of using Instagram and vulnerability to its mental health effects. Table A8 shows no significant differences in these characteristics between Instagram users and non-users in the pre-period. This lack of difference suggests that self-selection bias is unlikely to be a major concern in this analysis, as there is no evidence that individuals who chose to use Instagram had systematically different baseline characteristics related to mental health vulnerability compared to non-users.

The treatment in my analysis is defined as having an Instagram account by 2016, with the platform's algorithm introduced in March 2016 and the survey on social media use conducted in October 2016. This timeline raises a potential concern: individuals reporting an Instagram account in the 2016 survey may have been influenced by the introduction of the algorithm itself, potentially confounding the analysis. However, I argue that this is unlikely to significantly bias my findings. The relatively short window between the rollout

of the algorithm and the survey makes it improbable that the algorithm served as the primary motivation for most users to join Instagram. Broader social trends in platform adoption during this period likely played a much larger role, and including users from 2016 increases the sample size and improves the precision of my estimates, which is particularly important given the longitudinal nature of the survey. Nonetheless, to ensure the robustness of my treatment definition, I conducted additional analyses using a more restrictive approach, considering only individuals who created an Instagram account before 2016, meaning by October 2015, when the previous survey took place. Table A9 and Figure A9 show that the results remain highly consistent across the specifications, suggesting that the inclusion of 2016 users does not materially influence the findings. Furthermore, Table A10 highlights that, despite small statistically significant differences in variables such as the level of urbanization and education, individuals who joined Instagram before the introduction of the algorithm (before October 2015) and those who joined a few months after (October 2016) are largely comparable in their demographic and socioeconomic characteristics. While the 2016 group appears to live in slightly less urbanized areas and has a marginally higher level of education, these differences are minor. Crucially, mental health indicators, such as the general index of poor mental health and related measures, show no significant differences between the two groups, reinforcing the robustness of the treatment definition.

Additionally, I show that the results are not driven by the potential deterioration in mental health outcomes during the COVID-19 pandemic. Quarantines and lockdowns are known to induce states of isolation that are psychologically distressing and challenging for anyone experiencing them (Jiloha (2020)). Young people, already at higher risk of developing mental health issues compared to adults (Deighton et al. (2019)), may be especially susceptible to the adverse effects of such isolation. Factors like school closures, disrupted physical activity, and limited social interaction exacerbate these effects (Wang et al. (2020)). To address this, I exclude data from 2020 and 2021 in Table A11, showing the results from equation 1 without the pandemic period. The findings remain qualitatively consistent with the baseline results, indicating that pandemic-related mental health impacts do not drive the observed effects, thus reinforcing the robustness of the conclusions.

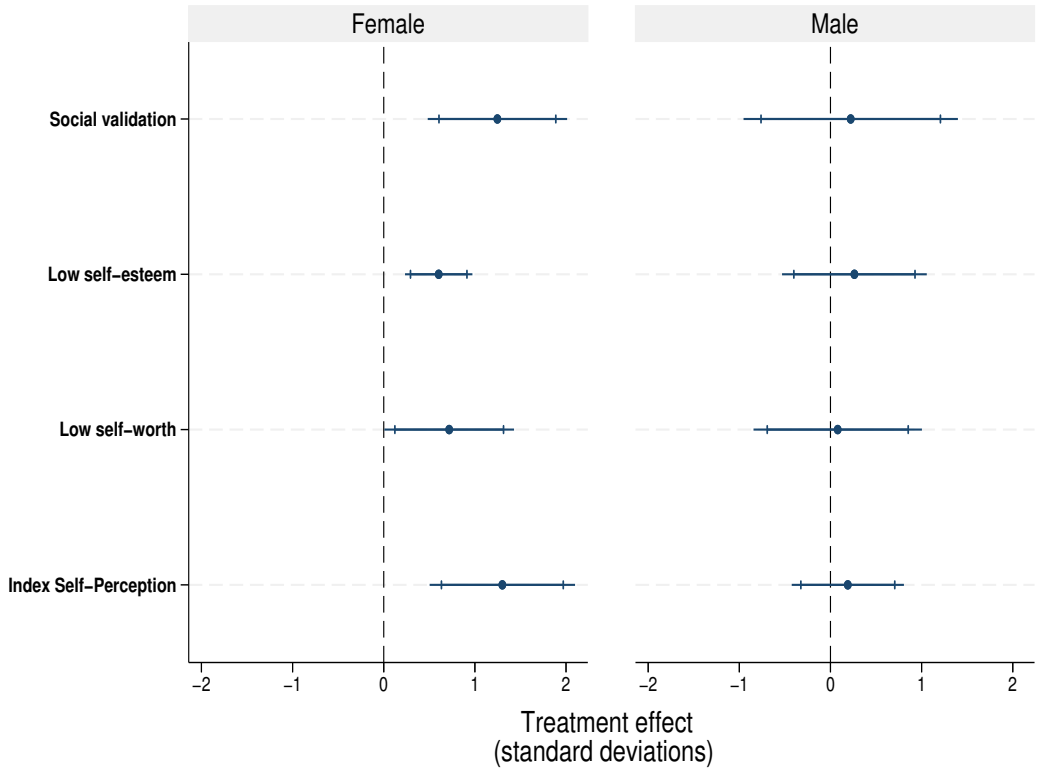
6 Mechanisms

Social media makes it easier for people to compare themselves to members of their social networks, and this comparisons could be detrimental to users' self-esteem and, consequently, mental health (Vogel et al. (2014); Bolognini et al. (1996)). With the introduction of the algorithm on Instagram, content is increasingly skewed toward posts that receive high engagement. This shift encourages individuals to seek similar recognition from others, reinforcing the importance of social validation. Social validation refers to the need for external approval, where one's sense of self-worth becomes increasingly dependent on the recognition of others.[9] This desire for validation often triggers social comparisons, which can lead to feelings of inadequacy, jealousy, or inferiority when individuals perceive themselves as not measuring up. The algorithmic nature of platforms like Instagram may therefore amplify social comparison behavior. This mechanism suggests that the algorithm encourages users to seek social validation, and as they compare themselves to others, they are more prone to feelings of inadequacy, leading to negative effects on self-esteem and self-worth. Figure 4 shows that the introduction of the algorithm increased the need for social validation and had a negative effect on self-esteem and self-worth (low self-esteem and low self-worth increase). These results are statistically significant only among female teenagers.

Moreover, the effect of the algorithm may be particularly pronounced for individuals who perceive themselves as unfavorable when compared to their peers. Figure 5 shows that the introduction of the algorithm on Instagram affected more severely the mental health of teenagers who might be more likely to be affected by unfavorable social comparisons. The figure displays estimates of the coefficient for the interaction between my treatment indicator and some moderators, with my index of poor mental health as the outcome variable. Specifically, I examine the following pre-period personality traits: jealousy of others' good fortune, avoidance of being the center of attention, reluctance to spend time on others, and a tendency to feel tense or unsettled. Moreover, I built an index of social comparisons based on these variables and consider, as an additional moderator, an indicator set to one if an individual social comparison index is above the median. These personality

[9] Go to https://www.psychology-lexicon.com/cms/glossary/52-glossary-s/24775-social-validation.html

Figure 4: Effects of the Instagram's algorithm on self-validation, self-esteem and self-worth

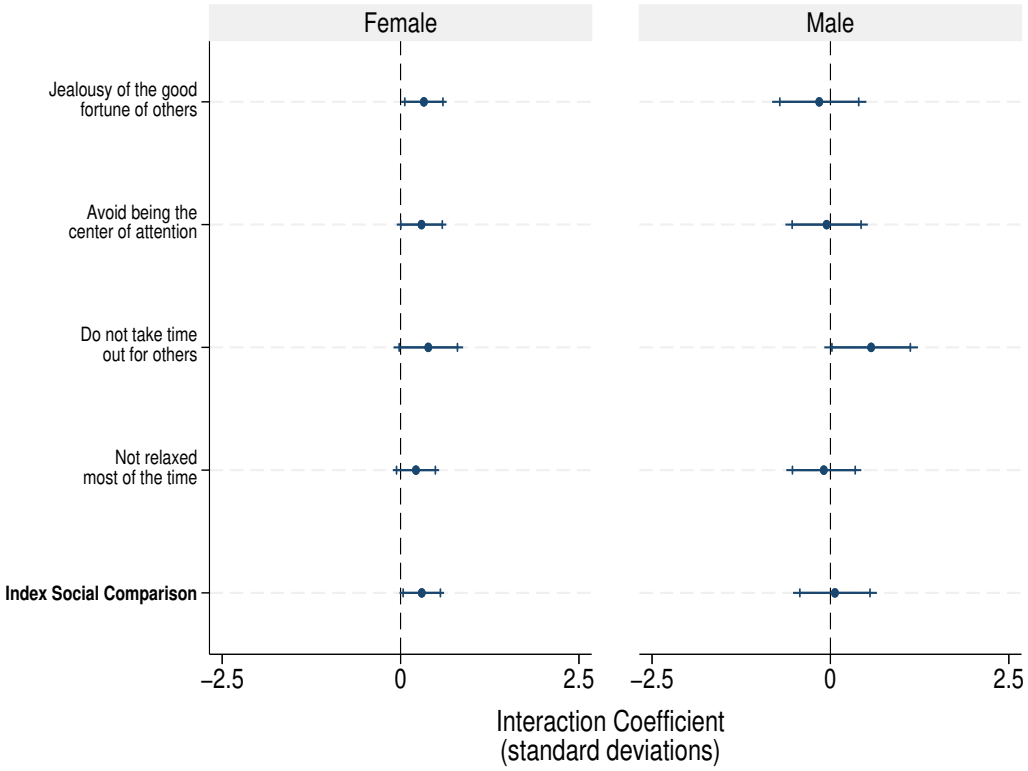


Notes. This Figure explores the effects of the introduction of the Instagram's algorithm on individuals' need for social validation, low self-esteem, low self-worth and index of the outcomes 1 through 3. All outcomes are standardized so that, in the preperiod, they have a mean of zero and a standard deviation of one. My controls consist of net income, level of education, housing situation and level of urbanization. For a detailed description of the outcome, treatment, and control variables, see Table A16. The bars represent 95 and 90 percent confidence intervals. Standard errors are clustered at the individual level.

traits, measured in the pre-treatment period, help identify individuals who may be more prone to negatively compare themselves to others. Traits such as jealousy of others' good fortune, discomfort with social attention, reluctance to engage socially, and a tendency toward anxiety or tension suggest a higher likelihood of engaging in social comparison in a self-critical way. Individuals with these characteristics are often more sensitive to how they measure up against others and may interpret others' successes, attractiveness, or social connections as a reflection of their own perceived shortcomings. For female teenager, all point estimates are positive, with a statistically significant effect observed for the overall index, jealousy of others' good fortune, avoidance of being the center of attention and reluctance to spend time on others. Aligned with the social comparison mechanism, the introduction of the algorithm on Instagram appears to have particularly adverse effects on

the mental health of individuals likely to perceive themselves as comparing unfavorably to their peers, because of some personality traits.

Figure 5: Heterogeneous effects as evidence of negative social comparisons



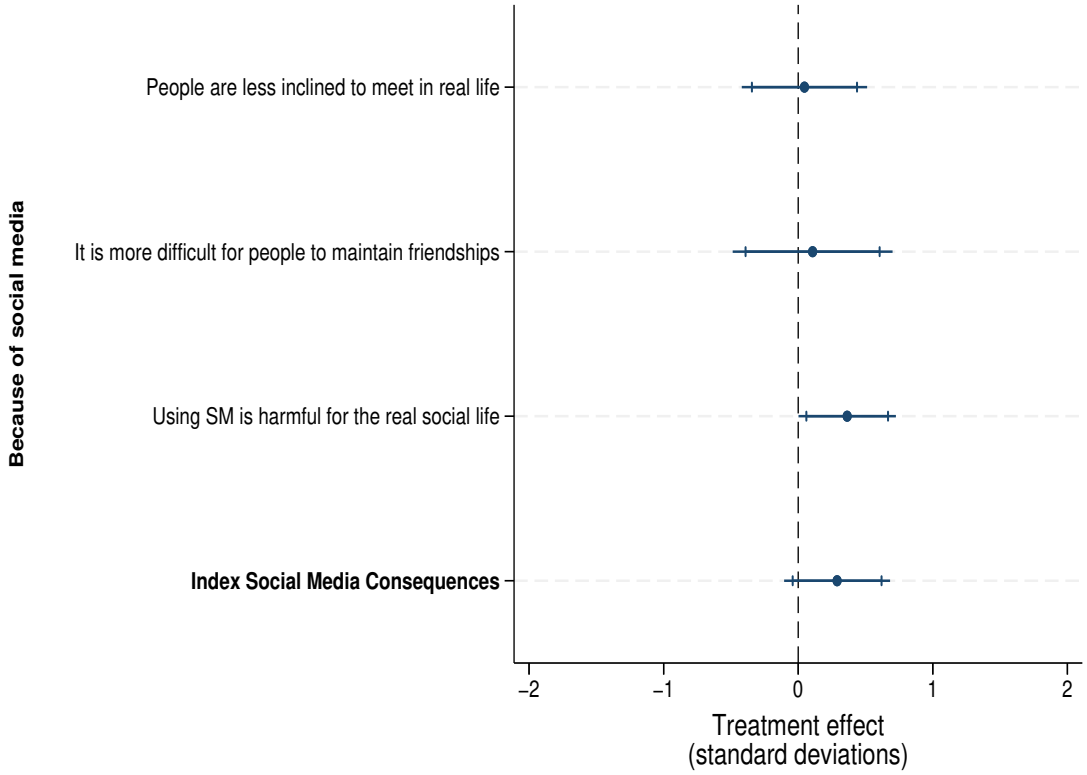
Notes. This figure investigates the mechanisms through which Instagram's algorithm impacts mental health. It presents estimates from a version of equation 1 where my treatment indicator interacts with indicators for specific pre-period personality traits. The outcome variable is my index of poor mental health. The index is standardized so that, in the pre-period, it has a mean of zero and a standard deviation of one. All estimates are obtained using my preferred specification, namely the one including controls and individual and time fixed effects. My controls consist of net income, level of education, housing situation and level of urbanization. For a detailed description of the outcome, treatment, and control variables, see Table A16. The bars represent 95 and 90 percent confidence intervals. Standard errors are clustered at the individual level.

The social media literature highlights a paradox: while social media generates significant consumer surplus, it also has negative effects on well-being. Bursztyn et al. (2023) conduct an online experiment to measure consumer welfare, accounting for both network effects and consumption spillovers to non-users. Their findings reveal what they term a "social media trap," where users experience substantial personal benefits from social media but also suffer from its negative welfare impacts. This trap reflects a situation in which users would prefer the platforms not to exist but struggle to coordinate an exit from them. A key driver behind this paradox is the fear of missing out (FoMO), a pervasive apprehension that others may be enjoying rewarding experiences in which one is not participating.

FoMO fuels a strong desire to stay constantly connected to others' activities (Przybylski et al. (2013)), reinforcing users' dependency on social media despite its detrimental effects on well-being. Moreover, evidence suggests that social comparison is associated with higher levels of FoMO (Reer et al. (2019); Burnell et al. (2019)). Indeed, people who tend to compare themselves with others more frequently are likely to expose themselves more often to the information others share about themselves, including positively biased information about their recent rewarding experiences and activities. As a result, they might more often conclude that others are doing better or having more rewarding experiences, which is a central aspect of FoMO (Przybylski et al. (2013)). I provide indirect evidence supporting the negative social comparison mechanism by demonstrating the existence of a "social media trap" within my context. Specifically, in Figure 6, I assess the impact of Instagram's algorithm introduction on teenagers' perceptions of social media's effects. The results show that teenagers' views on social media's impact on relationships remain consistent before and after the algorithm's introduction. However, they perceive that social media, following the algorithm's introduction, has a detrimental effect on their offline social lives, despite continuing to use the platform. This persistence highlights the "social media trap", where teenagers recognize the negative impact yet feel compelled to stay engaged. The persistence of the "social media trap" is further highlighted in Figure A11, which depicts exit rates from social media platforms before and after the introduction of Instagram's algorithm. The figure reveals no significant change in the likelihood of users exiting social media platforms following the algorithm's introduction. This suggests that, despite experiencing the algorithm's negative impacts on their offline lives, users remain either unwilling or unable to disengage from the platform en masse. When combined with the findings in Figure 6, which demonstrate that teenagers increasingly perceive social media as harmful to their offline relationships post-algorithm, these results provide compelling evidence supporting the existence of the "social media trap."

Some scholars argue that social media use can disrupt concentration, impair the ability to focus, and cause anxiety (Paul et al. (2012); Meier et al. (2016)). The rapid rise of online social networks has led some to suggest that excessive use of these platforms can be addictive for certain individuals (Kuss and Griffiths (2011)). The introduction of algorithms, which display content tailored to users' interests, could lead to prolonged

Figure 6: Effects of the Instagram's algorithm on the perceived consequences of social media



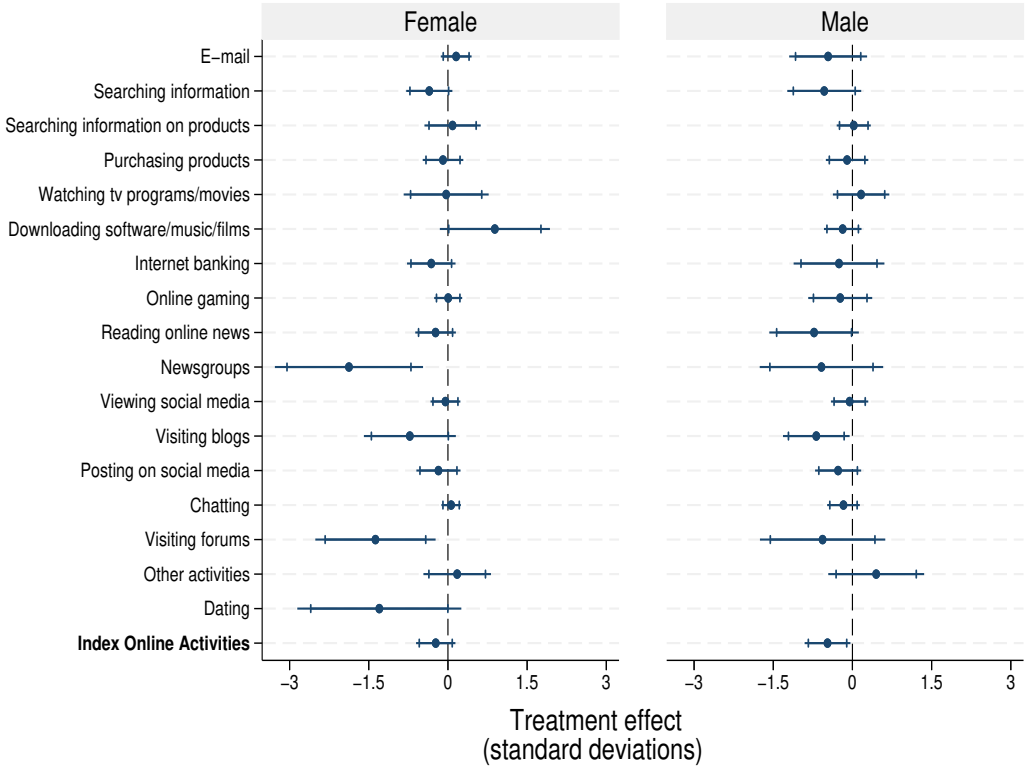
Notes. This Figure explores the effects of the introduction of the Instagram's algorithm on individuals' perceptions about the consequences of using social media. All outcomes are standardized so that, in the preperiod, they have a mean of zero and a standard deviation of one. All estimates are obtained using my preferred specification, namely the one including controls and individual and time fixed effects. My controls consist of net income, level of education, housing situation and level of urbanization. For a detailed description of the outcome, treatment, and control variables, see Table A16. The bars represent 95 and 90 percent confidence intervals. Standard errors are clustered at the individual level.

engagement on the platform (mindless scrolling), potentially resulting in overuse and addiction. To study this mechanism, I employ data on time spent on the Internet via different devices and time spent on various online activities, including social media. Figure A12 shows that the introduction of the Instagram's algorithm did not have a significant effect on average internet usage, across devices and by gender. Moreover, Figure 7 reveals that the time allocated to various online activities by both males and females after 2016 shows no significant increase, with the notable exception of a rise among female teenagers in the time spent downloading software, music, and films. This trend might be influenced by Instagram's personalized content and advertising strategies, which often promote apps and tools related to media consumption, entertainment, and personal interests. Such promotions likely resonate more strongly with young women, reflecting broader patterns in

user engagement and preferences. At the same time, female teenagers appear to spend slightly less time on certain activities, such as newsgroups, blogs, forums, and dating websites, while boys spend less time reading online news and visiting blogs. The observed decline in engagement with dating platforms among girls is particularly intriguing and may stem from several interconnected factors. First, the findings suggest that Instagram's algorithm negatively affects self-esteem and heightens social comparison, which could lead to reduced confidence in one's appearance and greater hesitancy to participate in platforms where physical looks are emphasized, such as dating apps. Second, Instagram itself increasingly serves as a space for informal dating interactions through features like personalized content and direct messaging, potentially diminishing the reliance on dedicated dating platforms. Notably, there is no observable effect on the total time spent viewing and posting on social media. These variables capture overall social media usage without distinguishing between platforms, so it is unclear whether individuals shifted time from other social media platforms to Instagram. To address this, Figure A13 illustrates the effect of the algorithm's introduction on mental health by dividing the sample into subgroups based on the hours spent on social media. The magnitude increases with the hours spent on social media, yet the effect remains significant across all groups. It is important to note that the reported hours spent online are self-reported, introducing the potential for measurement error. This issue is particularly relevant given the nature of social media usage: interactions are often fragmented and sporadic, with individuals frequently picking up their smartphones for short intervals without consciously tracking the cumulative time spent. As a result, it becomes difficult for users to accurately estimate their total hours online, which could lead to underreporting. This limitation might partially explain the absence of any observable effect on the total time spent viewing and posting on social media, as such inaccuracies could mask potential shifts in usage patterns.

Next, there might be concern that the introduction of the algorithm on Instagram influenced the stigma associated with mental illness and that my results do not reflect an increase in the prevalence of mental illness per se, but rather an increase in the willingness to talk about it. The idea is that with the introduction of the algorithm, Instagram users might have been more exposed to content that normalizes mental illness than users of other social networks. If Instagram's algorithm made people more comfortable talking about

Figure 7: Effects of the Instagram's algorithm on internet activities



Notes. This Figure explores the effects of the introduction of the Instagram's algorithm on average hours spent on several online activities. All outcomes are standardized so that, in the preperiod, they have a mean of zero and a standard deviation of one. All estimates are obtained using my preferred specification, namely the one including controls and individual and time fixed effects. My controls consist of net income, level of education, housing situation and level of urbanization. For a detailed description of the outcome, treatment, and control variables, see Table A16. The bars represent 95 and 90 percent confidence intervals. Standard errors are clustered at the individual level.

mental illness, I would expect to observe fewer missing responses after the introduction of it. Consistent with the effects being due to higher prevalence of mental illness rather than stigma, Table A12 shows that the prevalence of missing responses was not affected by the introduction of the Instagram algorithm. Furthermore, if one believes that the algorithm's effect on mental health stems from an increased willingness to report symptoms, it would be reasonable to expect that the algorithm's introduction might lead to a greater increase in reporting of less severe mental health conditions compared to more severe ones. Accordingly, in Table A13, I present the results of a separate analysis, using as outcome variable the number of missing values for each individual mental health questions. The findings again suggest that the observed effects are due to an actual increase in the prevalence of mental illness, rather than a heightened tendency to report mental health issues. Indeed, no significant effect is observed for any specific variable, indicating that the results are

not driven by teenagers reporting mental health struggles more frequently following the introduction of the algorithm. Finally, Figure A14 shows the share of Dutch teenagers and young people aged 15 to 25 who were prescribed antidepressants, anxiolytics, and antipsychotics between 2014 and 2021.[10] In particular, Figure A14 shows that, starting in 2016, the share of girls prescribed antidepressants increases. This evidence aligns with previous findings, suggesting that my result corresponds to a higher prevalence of mental illness and is not simply the result of increased reporting or reduced stigma associated with these disorders. In fact, medication prescriptions, being under the purview of medical professionals, should be the result of a medical evaluation and therefore less prone to reporting bias. Furthermore, in Table A14 I show that the introduction of the algorithm on Instagram did not affect the reporting of other stigmatized conditions, such as illegal substance use. If stigma reduction was indeed the driving force behind my results, it would be surprising not to find similar effects on other stigmatized behaviors and conditions.

Finally, social media may have caused teenagers to engage or not engage in a number of other behaviors that have some effect on mental health, and this may be amplified by the introduction of the algorithm. Tables A14, A15 and A3 show that there are no effects on drug use, drinking behaviors and physical activity practice.

7 External Validity

In this section, I provide evidence suggesting that the identified effects are not unique to the Netherlands and likely have broader applicability. To support this, I use repeated cross-sectional data from the U.S. around the time Instagram introduced its algorithm. Specifically, I analyze two survey waves conducted by the Pew Research Center, which include questions on social media use and perceived effects of these platforms.[11] Unlike the longitudinal design outlined in Section 4, this survey is cross-sectional, so it does not allow for tracking individuals over time. However, I can compare Instagram users before and after the algorithm's introduction with users of other social media platforms, as the survey collects data on platform use.

[10] I use data from Statistics Netherlands.

[11] I use the Teen Relationship Survey Pretest (2014-2015) and the Teen Survey (2018), focusing on individuals born between 1997 and 2005

Figure A15 demonstrates that, after the algorithm's introduction, individuals were more likely to report that social media made them feel worse about their own lives, increased pressure to post content that portrays them positively, and heightened the importance of receiving likes and comments. The first finding aligns with my previous results, indicating that the algorithm's introduction on Instagram has a negative effect on teenager mental health and well-being. The latter effects provide support for the negative social comparisons mechanism. Specifically, the algorithmic feed seems to increase teenagers' pressure to seek social validation through likes and comments and to post content that enhances their appearance, striving to "measure up" to others to whom they compare themselves upwardly.

8 Conclusions

In 2023, over half the global population had a social media account, with the average user spending around two and a half hours daily on these platforms (We Are Social 2024). Since social media's rise in the mid-2000s, teenager and young adult mental health has shown noticeable declines. The evolution of social media platforms, especially through the adoption of algorithms, has introduced significant changes. Algorithms now curate and personalize content, which benefits some users while marginalizing others. Although research has explored social media's overall impact on mental health, specific platform features, such as algorithms, remain under-examined.

This paper examines one of the most transformative changes in social media, recommender algorithms, assessing their impact on users' mental health. Specifically, it provides quasi-experimental evidence on the effect of Instagram's 2016 algorithm introduction by analyzing data from the Dutch Longitudinal Internet Studies for the Social Sciences (LISS) panel. Using a difference-in-differences approach, I evaluate the impact of this algorithmic shift on teenagers' mental health. Findings indicate a significant negative effect, with the poor mental health index for teenagers worsening by an estimated 0.394 standard deviation units after the algorithm's implementation. Importantly, this effect is not attributable to reduced stigma around mental health issues or increased likelihood of reporting such conditions.

The analysis suggests that the primary mechanism behind this negative impact is intensified negative social comparison. Instagram's algorithm prioritizes high-engagement posts, prompting users to seek external validation and social approval. This need for social validation, where self-worth increasingly hinges on others' approval, drives frequent comparisons, often leaving users feeling inadequate, envious, or inferior. As a result, the algorithm amplifies social comparison behaviors, creating a cycle where users become increasingly dependent on external validation, which, in turn, harms self-esteem and self-worth. Moreover, aligning with the concept of the "social media trap" outlined by Bursztyn et al. (2023), I provide compelling evidence that users continue to engage with the platform despite recognizing its harmful effects on their offline lives. This paradoxical behavior is largely driven by fear of missing out (FoMO), which perpetuates their reliance on the platform. Consequently, even as users acknowledge the negative consequences of social media, they feel unable to disengage, thereby exacerbating the detrimental impacts on self-esteem and well-being.

The results align with the hypothesis that social media may contribute to the recent decline in young people's mental health. Moreover, these findings underscore the profound mental health implications of algorithmic content curation on social media platforms. The paper calls for policymakers and social media companies to consider the mental health impacts of algorithmic design, particularly for vulnerable populations such as teenagers.

References

Agung, N. F. A. and Darma, G. S. (2019). Opportunities and challenges of instagram algorithm in improving competitive advantage. *International Journal of Innovative Science and Research Technology*, 4(1):743–747.

Allcott, H., Braghieri, L., Eichmeyer, S., and Gentzkow, M. (2020). The welfare effects of social media. *American Economic Review*, 110(3):629–676.

Allcott, H., Gentzkow, M., and Song, L. (2022). Digital addiction. *American Economic Review*, 112(7):2424–2463.

Anderson, D. M., Cesur, R., and Tekin, E. (2015). Youth depression and future criminal behavior. *Economic Inquiry*, 53(1):294–317.

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool, and early training projects. *Journal of the American statistical Association*, 103(484):1481–1495.

Arenas-Arroyo, E., Fernández-Kranz, D., and Nollenberger, N. (2022). High speed internet and the widening gender gap in adolescent mental health: evidence from hospital records. Working paper, IZA Discussion Papers, No. 15728.

Biasi, B., Dahl, M. S., and Moser, P. (2021). Career effects of mental health. Working paper, National Bureau of Economic Research, No. w29031.

Bolognini, M., Plancherel, B., Bettschart, W., and Halfon, O. (1996). Self-esteem and mental health in early adolescence: Development and gender differences. *Journal of adolescence*, 19(3):233–245.

Braghieri, L., Levy, R., and Makarin, A. (2022). Social media and mental health. *American Economic Review*, 112(11):3660–3693.

Burnell, K., George, M. J., Vollet, J. W., Ehrenreich, S. E., and Underwood, M. K. (2019). Passive social networking site use and well-being: The mediating roles of social comparison and the fear of missing out.

Bursztyn, L., Handel, B. R., Jimenez, R., and Roth, C. (2023). When product markets become collective traps: The case of social media. Working paper, National Bureau of Economic Research, No. w31771.

Castellacci, F. and Tveito, V. (2018). Internet use and well-being: A survey and a theoretical framework. *Research policy*, 47(1):308–325.

Currie, J. and Stabile, M. (2006). Child mental health and human capital accumulation: the case of adhd. *Journal of health economics*, 25(6):1094–1118.

De Chaisemartin, C. and d'Haultfoeuille, X. (2024). Difference-in-differences estimators of intertemporal treatment effects. *Review of Economics and Statistics*, pages 1–45.

Deighton, J., Lereya, S. T., Casey, P., Patalay, P., Humphrey, N., and Wolpert, M. (2019). Prevalence of mental health problems in schools: poverty and other risk factors among 28 000 adolescents in england. *The British Journal of Psychiatry*, 215(3):565–567.

Donati, D., Durante, R., Sobbrio, F., and Zejcirovic, D. (2022). Lost in the net? broadband internet and youth mental health. *Broadband Internet and Youth Mental Health (March 2022)*.

Germano, F., Gómez, V., and Sobbrio, F. (2022). Ranking for engagement: How social media algorithms fuel misinformation and polarization. Working paper, CESifo Working Paper No. 10011.

Golberstein, E., Gonzales, G., and Meara, E. (2019). How do economic downturns affect the mental health of children? evidence from the national health interview survey. *Health economics*, 28(8):955–970.

Golin, M. (2022). The effect of broadband internet on the gender gap in mental health: Evidence from germany. *Health Economics*, 31:6–21.

Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., Crespo-Tenorio, A., Dimmery, D., Freelon, D., Gentzkow, M., et al. (2023). How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science*, 381(6656):398–404.

Haushofer, J. and Fehr, E. (2014). On the psychology of poverty. *science*, 344(6186):862–867.

Haushofer, J. and Shapiro, J. (2016). The short-term impact of unconditional cash transfers to the poor: experimental evidence from kenya. *The Quarterly Journal of Economics*, 131(4):1973–2042.

Hawkley, L. C. and Cacioppo, J. T. (2010). Loneliness matters: A theoretical and empirical review of consequences and mechanisms. *Annals of behavioral medicine*, 40(2):218–227.

Huszár, F., Ktena, S. I., O'Brien, C., Belli, L., Schlaikjer, A., and Hardt, M. (2022). Algorithmic amplification of politics on twitter. *Proceedings of the National Academy of Sciences*, 119(1):e2025334119.

Jiloha, R. (2020). Covid-19 and mental health. *Epidemiology International (E-ISSN: 2455-7048)*, 5(1):7–9.

Kuss, D. J. and Griffiths, M. D. (2011). Online social networking and addiction—a review of the psychological literature. *International journal of environmental research and public health*, 8(9):3528–3552.

Levy, R. (2021). Social media, news consumption, and polarization: Evidence from a field experiment. *American economic review*, 111(3):831–870.

Meier, A., Reinecke, L., and Meltzer, C. E. (2016). "facebocrastination"? predictors of using facebook for procrastination and its effects on students' well-being. *Computers in Human Behavior*, 64:65–76.

Mosquera, R., Odunowo, M., McNamara, T., Guo, X., and Petrie, R. (2020). The economic effects of facebook. *Experimental Economics*, 23:575–602.

Mphaphuli, L. K. (2023). The impact of dysfunctional families on the mental health of children. In *Parenting in Modern Societies*. IntechOpen.

Nyhan, B., Settle, J., Thorson, E., Wojcieszak, M., Barberá, P., Chen, A. Y., Allcott, H., Brown, T., Crespo-Tenorio, A., Dimmery, D., et al. (2023). Like-minded sources on facebook are prevalent but not polarizing. *Nature*, 620(7972):137–144.

OECD (2022). Health at a glance: Europe 2020: State of health in the eu cycle.

Patel, V., Chisholm, D., Parikh, R., Charlson, F. J., Degenhardt, L., Dua, T., Ferrari, A. J., Hyman, S., Laxminarayan, R., Levin, C., et al. (2016). Addressing the burden of mental, neurological, and substance use disorders: key messages from disease control priorities. *The Lancet*, 387(10028):1672–1685.

Paul, J. A., Baker, H. M., and Cochran, J. D. (2012). Effect of online social networking on student academic performance. *Computers in human behavior*, 28(6):2117–2127.

Paul, K. I. and Moser, K. (2009). Unemployment impairs mental health: Meta-analyses. *Journal of Vocational behavior*, 74(3):264–282.

Persson, P. and Rossin-Slater, M. (2018). Family ruptures, stress, and the mental health of the next generation. *American economic review*, 108(4-5):1214–1252.

Przybylski, A. K., Murayama, K., DeHaan, C. R., and Gladwell, V. (2013). Motivational, emotional, and behavioral correlates of fear of missing out. *Computers in human behavior*, 29(4):1841–1848.

Reer, F., Tang, W. Y., and Quandt, T. (2019). Psychosocial well-being and social media engagement: The mediating roles of social comparison orientation and fear of missing out. *New Media & Society*, 21(7):1486–1505.

Ridley, M., Rao, G., Schilbach, F., and Patel, V. (2020). Poverty, depression, and anxiety: Causal evidence and mechanisms. *Science*, 370(6522):eaay0214.

Scherpenzeel, A. C. (2018). ""true" longitudinal and probability-based internet panels: Evidence from the netherlands. In *Social and behavioral research and the internet*, pages 77–104. Routledge.

Sela, Y., Zach, M., Amichay-Hamburger, Y., Mishali, M., and Omer, H. (2020). Family environment and problematic internet use among adolescents: The mediating roles of depression and fear of missing out. *Computers in Human Behavior*, 106:106226.

Thomas, P. A., Liu, H., and Umberson, D. (2017). Family relationships and well-being. *Innovation in aging*, 1(3):igx025.

Tineo, P., Bixter, M. T., Polanco-Roman, L., Grapin, S. L., Taveras, L., and Reyes-Portillo, J. (2024). The impact of acculturative stress on internalizing problems among racially and ethnically minoritized adolescents and young adults in the us: A systematic review and meta-analysis. *Social Science & Medicine*, page 117192.

Twenge, J. M. and Campbell, W. K. (2019). Media use is linked to lower psychological well-being: Evidence from three datasets. *Psychiatric Quarterly*, 90:311–331.

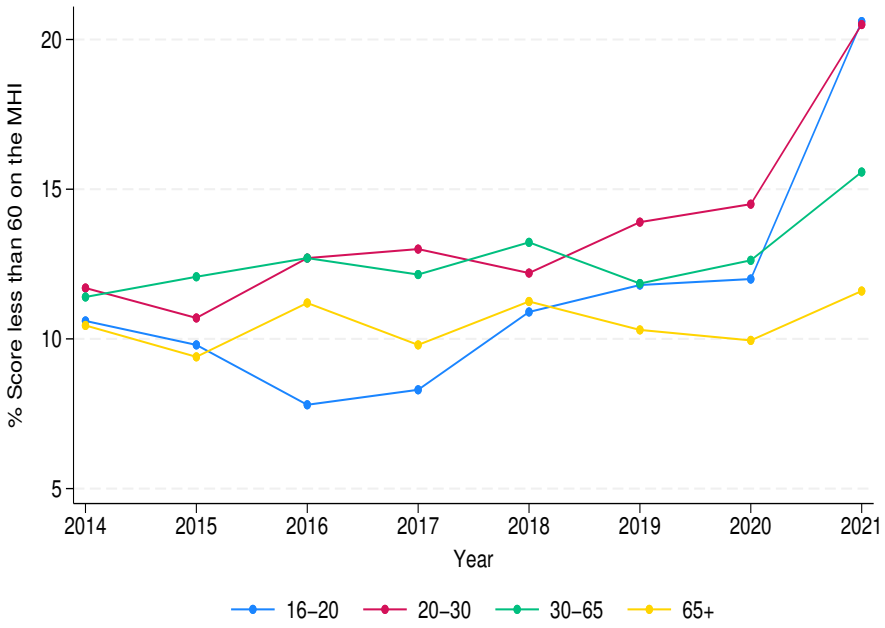
Vogel, E. A., Rose, J. P., Roberts, L. R., and Eckles, K. (2014). Social comparison, social media, and self-esteem. *Psychology of popular media culture*, 3(4):206.

Wang, G., Zhang, Y., Zhao, J., Zhang, J., and Jiang, F. (2020). Mitigate the effects of home confinement on children during the covid-19 outbreak. *The lancet*, 395(10228):945–947.

A Appendix

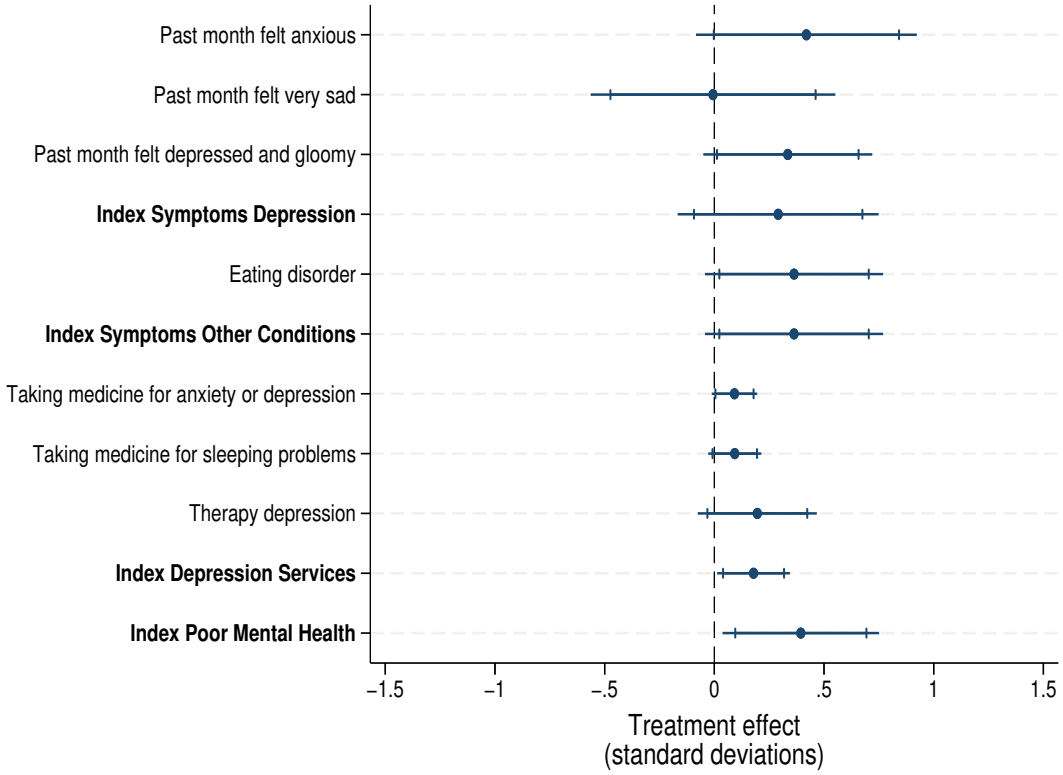
A.1 Additional Tables and Figures

Figure A1: Mental health trends in the Netherlands, 2014–2021



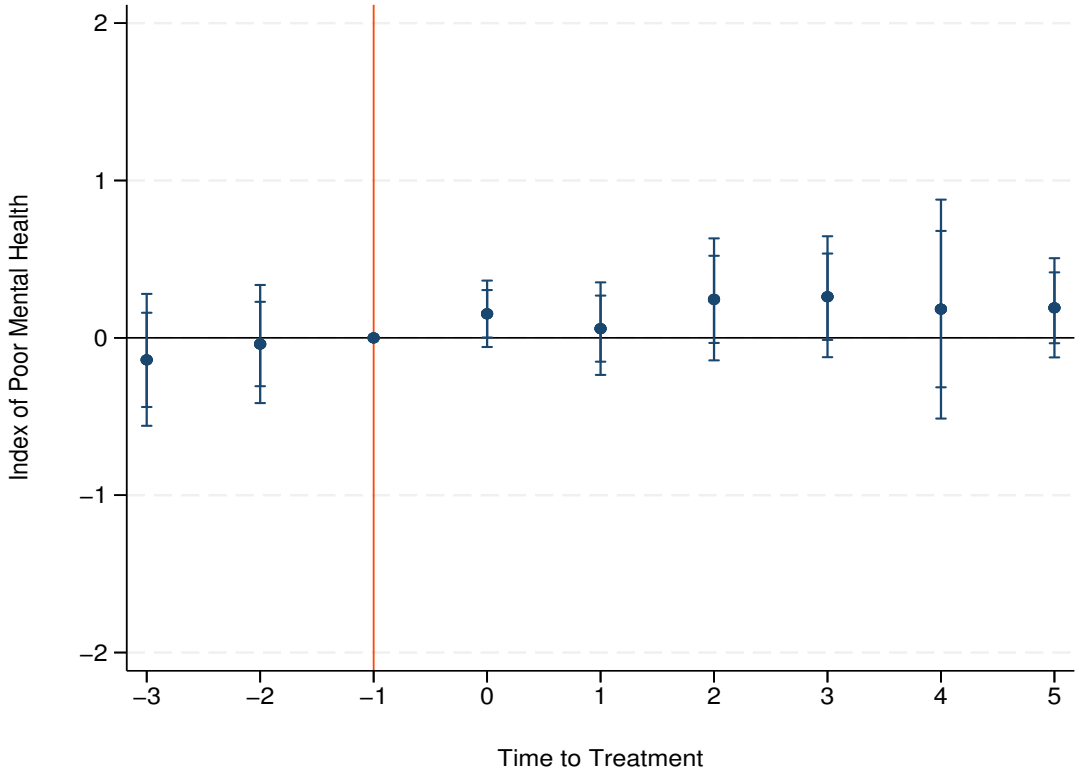
Notes. This Figure displays mental health trends in the Netherlands by age group in 2014–2021. The data come from the Health Survey conducted by Statistics Netherlands (Centraal Bureau voor de Statistiek). For a detailed description of the outcome, treatment, and control variables, see Table A16.

Figure A2: Effects of the introduction of the Instagram's algorithm on teenagers mental health



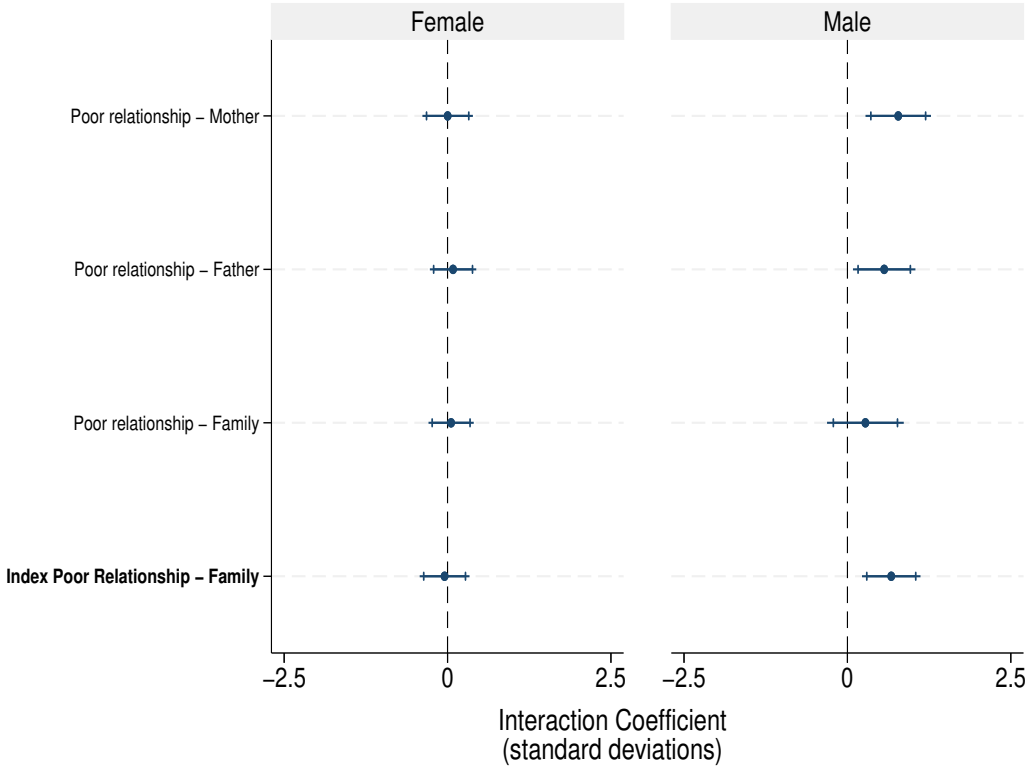
Notes. This Figure explores the effects of the introduction of the Instagram's algorithm on all my mental-health outcome variables and on the related indices. It display estimates of coefficient β from equation 1 using my preferred specification, namely the one including controls and individual and time fixed effects. The outcome variables are my overall index of poor mental health, the individual components of the index, and three subindices: the index of depression symptoms, the index of symptoms of other mental health conditions, and the index of depression services. All outcomes are standardized so that, in the preperiod, they have a mean of zero and a standard deviation of one. My controls consist of net income, level of education, housing situation and level of urbanization. For a detailed description of the outcome, treatment, and control variables, see Table A16. The bars represent 95 and 90 percent confidence intervals. Standard errors are clustered at the individual level.

Figure A3: Event study for the Index of Poor Mental Health - De Chaisemartin and d'Haultfoeuille (2024)



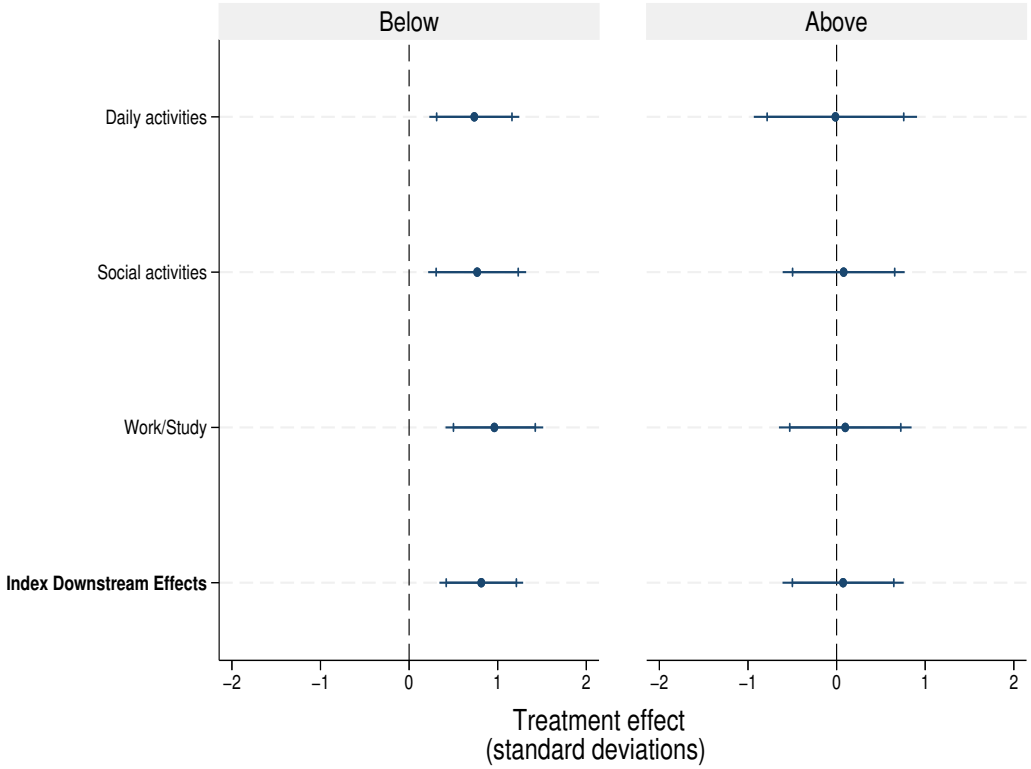
Notes. This Figure overlays the event-study plot using the robust estimator proposed by De Chaisemartin and d'Haultfoeuille (2024). The outcome variable is my index of poor mental health. The index is standardized so that, in the pre-period, it has a mean of zero and a standard deviation of one. For a detailed description of the outcome, treatment, and control variables, see Table A16. The bars represent 95 and 90 percent confidence intervals. Standard errors are clustered at the individual level.

Figure A4: Heterogenous effects - family characteristics



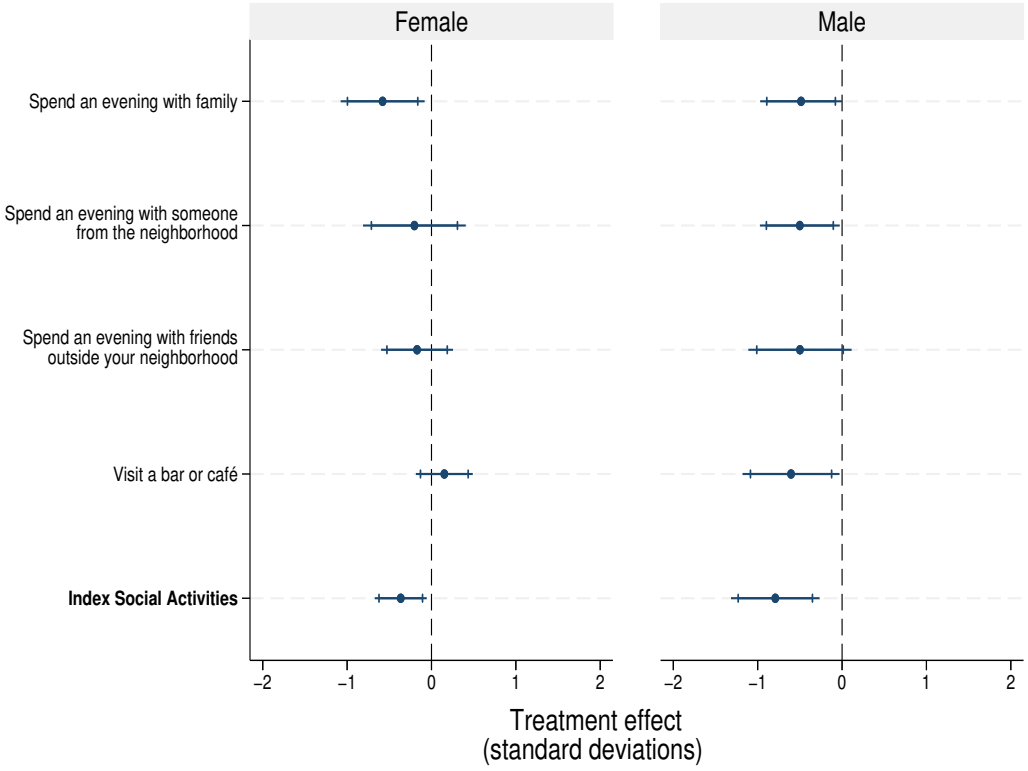
Notes. This Figure explores whether the effects of the introduction of the algorithm on Instagram on mental health are heterogeneous across a host of family characteristics. Specifically, it presents estimates from a version of equation 1 in which my treatment indicator is interacted with various moderators. The outcome variable is my index of poor mental health. The index is standardized so that, in the pre-period, it has a mean of zero and a standard deviation of one. All estimates are obtained using my preferred specification, namely the one including controls and individual and time fixed effects. My controls consist of net income, level of education, housing situation and level of urbanization. For a detailed description of the outcome, treatment, and control variables, see Table A16. The bars represent 95 and 90 percent confidence intervals. Standard errors are clustered at the individual level.

Figure A5: Downstream effects due to emotional problems



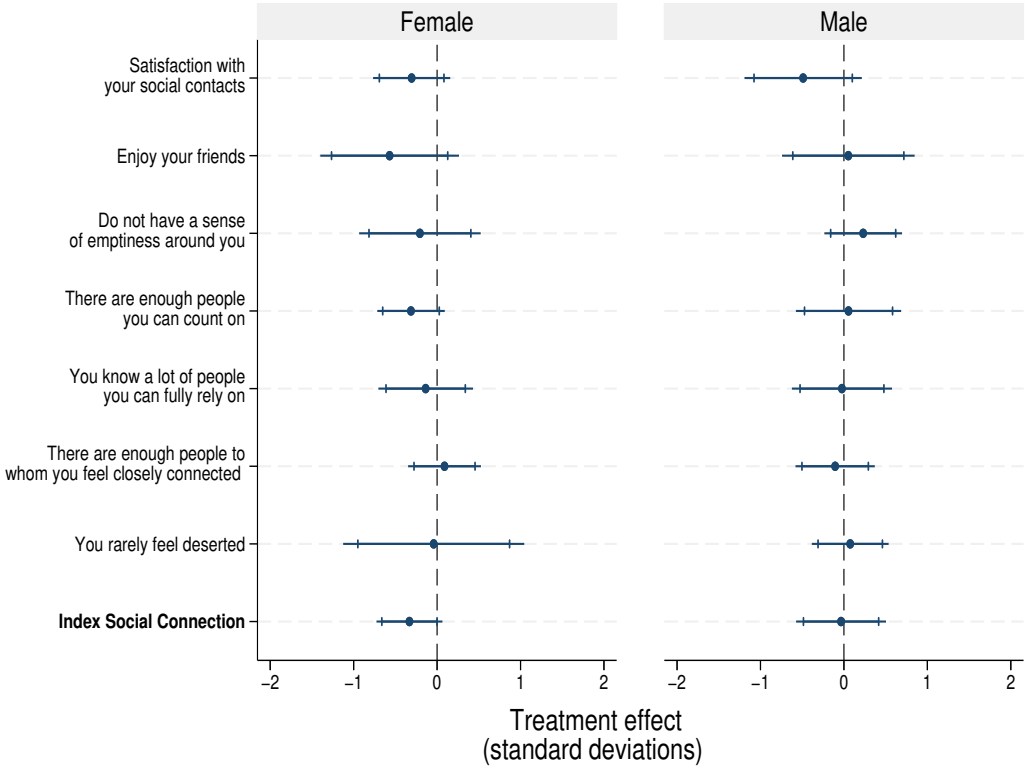
Notes. This figure examines the downstream effects of Instagram's algorithm on teenagers' ability to perform various activities, focusing on differences between those above and below the median of my index of poor physical health. It presents estimates of the coefficient β from equation 1 separately for these two groups, allowing me to assess the impact of emotional difficulties on different activities. Indeed, the outcome variables are answers to questions inquiring as to whether physical health or emotional problems hinder their daily activities, social activities, work and my index of downstream effects. All outcomes are standardized so that, in the pre-period, they have a mean of zero and a standard deviation of one. All estimates are obtained using my preferred specification, namely the one including controls and individual and time fixed effects. My controls consist of net income, level of education, housing situation and level of urbanization. For a detailed description of the outcome, treatment, and control variables, see Table A16. The bars represent 95 and 90 percent confidence intervals. Standard errors are clustered at the individual level.

Figure A6: Effects of the Instagram's algorithm on social activities



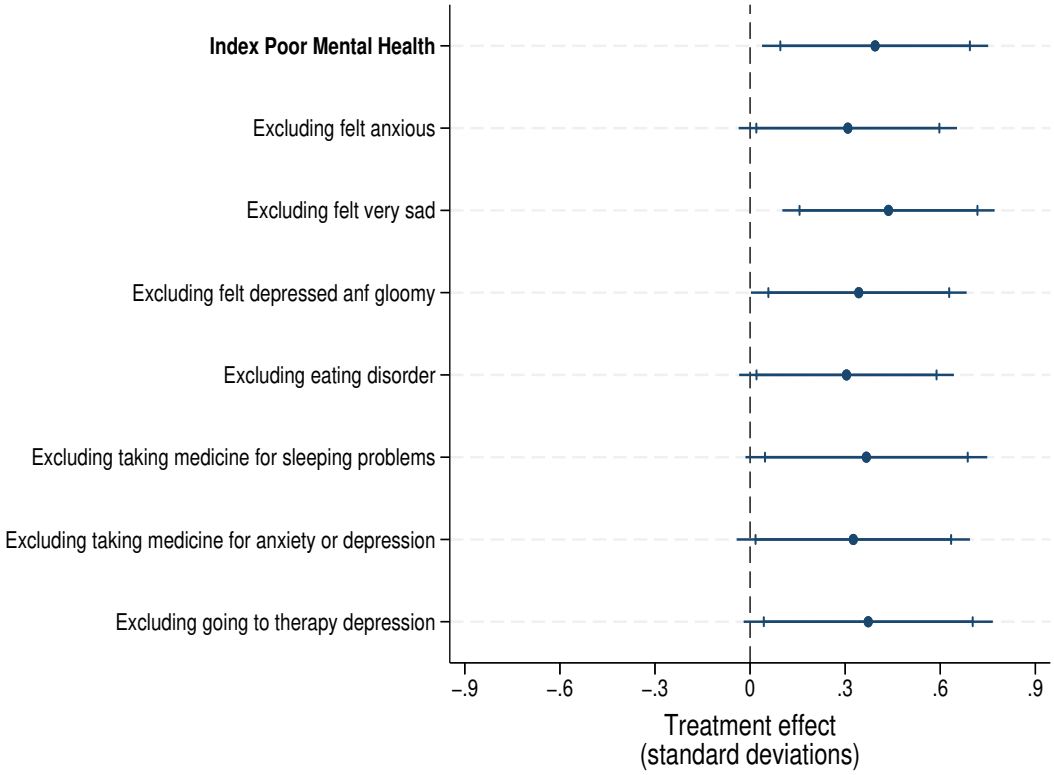
Notes. This Figure explores the effects of the introduction of the Instagram's algorithm on the patterns of individuals' engagement in different social activities. All outcomes are standardized so that, in the preperiod, they have a mean of zero and a standard deviation of one. All estimates are obtained using my preferred specification, namely the one including controls and individual and time fixed effects. My controls consist of net income, level of education, housing situation and level of urbanization. For a detailed description of the outcome, treatment, and control variables, see Table A16. The bars represent 95 and 90 percent confidence intervals. Standard errors are clustered at the individual level.

Figure A7: Effects of the Instagram's algorithm on social connection quality measures



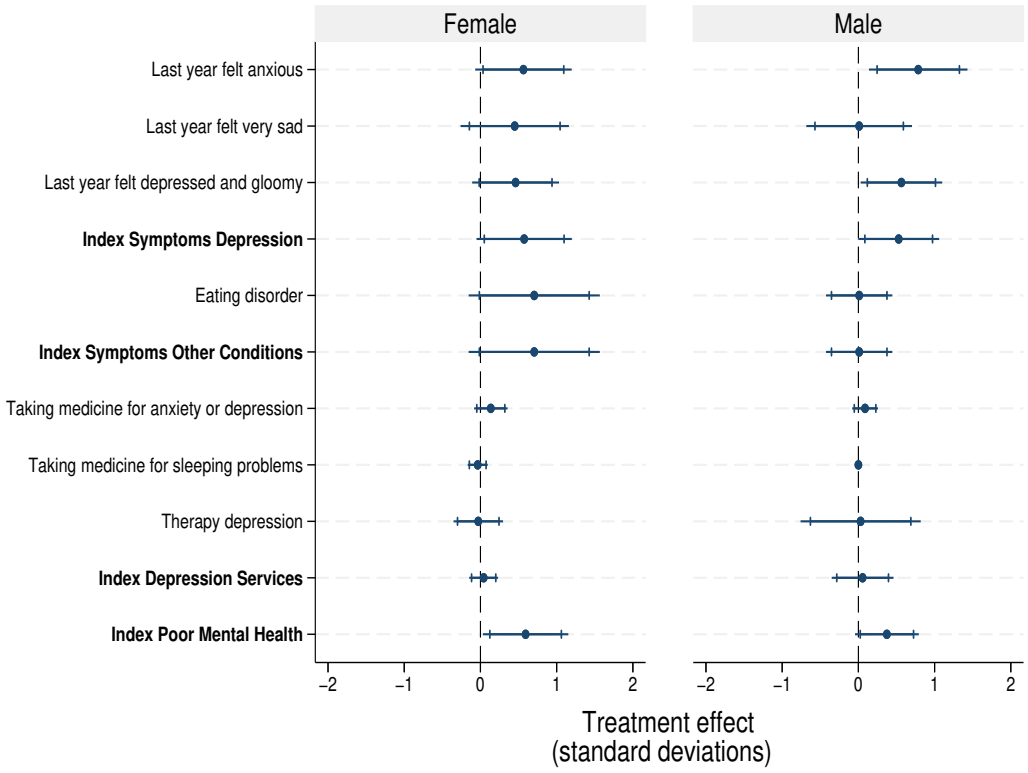
Notes. This Figure explores the effects of the introduction of the Instagram algorithm on on various measures of social connection quality. All variables are standardized so that, over the previous period, they have a mean of zero and a standard deviation of one. All estimates are obtained using my preferred specification, namely the one including controls and individual and time fixed effects. My controls consist of net income, level of education, housing situation and level of urbanization. For a detailed description of the outcome, treatment, and control variables, see Table A16. The bars represent 95 and 90 percent confidence intervals. Standard errors are clustered at the individual level.

Figure A8: Robustness to excluding each variable from the Index of Poor Mental Health



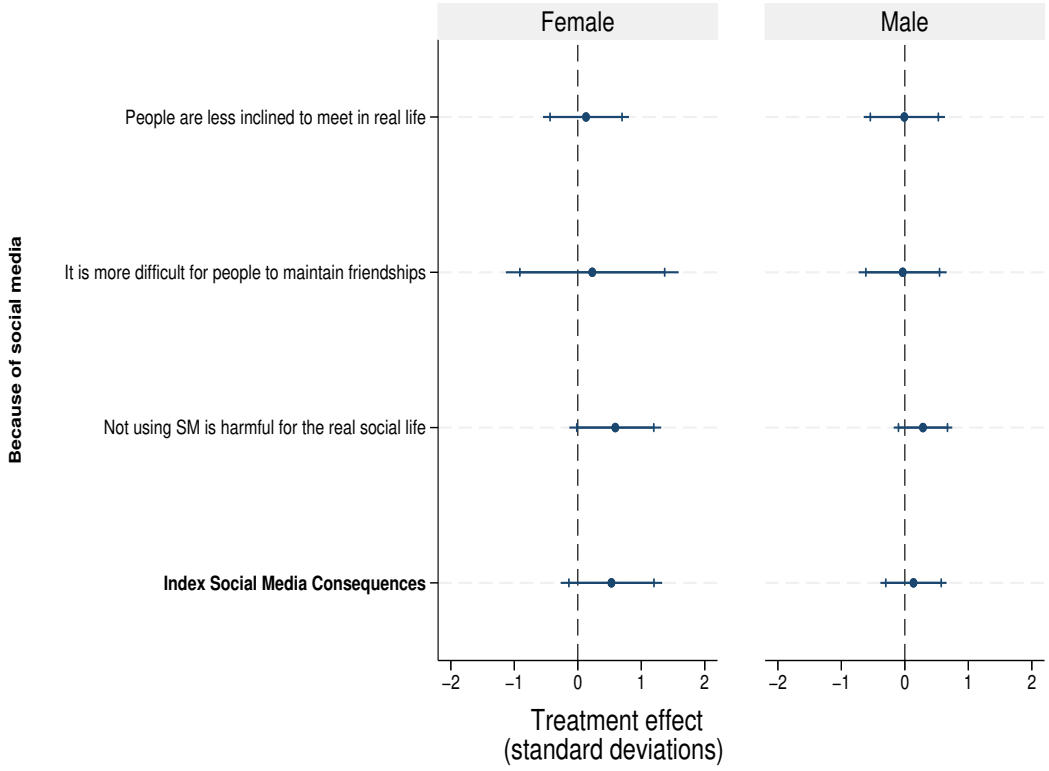
Notes. This Figure explores the robustness of my baseline results to excluding each individual variable from the construction of the index of poor mental health. It display estimates of coefficient β from equation 1 using my preferred specification, namely the one including controls and individual and time fixed effects. Each row excludes a different variable from the construction of the index. All outcomes are standardized so that, in the preperiod, they have a mean of zero and a standard deviation of one. For a detailed description of the outcome, treatment, and control variables, see Table A16. My controls consist of net income, level of education, housing situation and level of urbanization. Standard errors are clustered at the individual level.

Figure A9: Effects of the Instagram's algorithm on teenagers mental health by gender - with treatment defined as Instagram users until October 2015



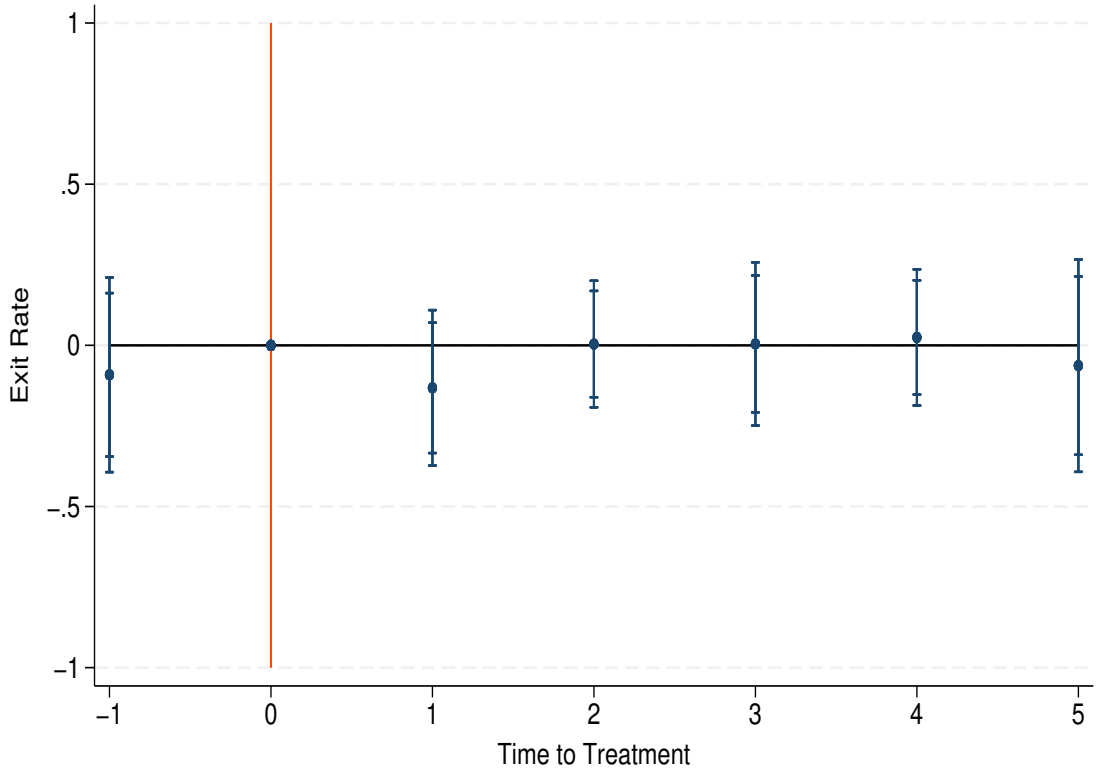
Notes. This Figure explores the effects of the introduction of the Instagram's algorithm on all my mental health outcome variables and on the related indices, by gender, and where the treatment is constructed as individuals who reported having an Instagram account by October 2015. It displays estimates of coefficient β from equation 1 using my preferred specification, namely the one including controls and individual and time fixed effects. The outcome variables are my overall index of poor mental health, the individual components of the index, and three subindices: the index of depression symptoms, the index of symptoms of other mental health conditions, and the index of depression services. All outcomes are standardized so that, in the preperiod, they have a mean of zero and a standard deviation of one. My controls consist of net income, level of education, housing situation and level of urbanization. For a detailed description of the outcome, treatment, and control variables, see Table A16. The bars represent 95 and 90 percent confidence intervals. Standard errors are clustered at the individual level.

Figure A10: Effects of the Instagram's algorithm on the perceived consequences of social media, by gender



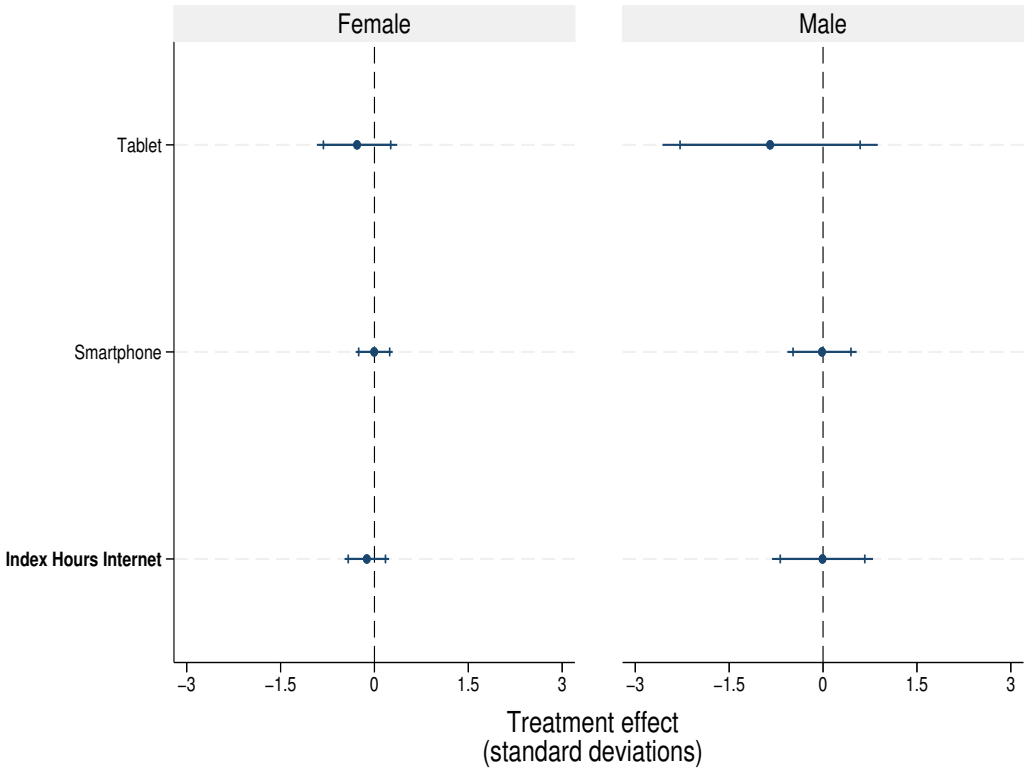
Notes. This Figure explores the effects of the introduction of the Instagram's algorithm on individuals' perceptions about the consequences of using social media. All outcomes are standardized so that, in the preperiod, they have a mean of zero and a standard deviation of one. All estimates are obtained using my preferred specification, namely the one including controls and individual and time fixed effects. My controls consist of net income, level of education, housing situation and level of urbanization. For a detailed description of the outcome, treatment, and control variables, see Table A16. The bars represent 95 and 90 percent confidence intervals. Standard errors are clustered at the individual level.

Figure A11: Exit rate from social media platforms



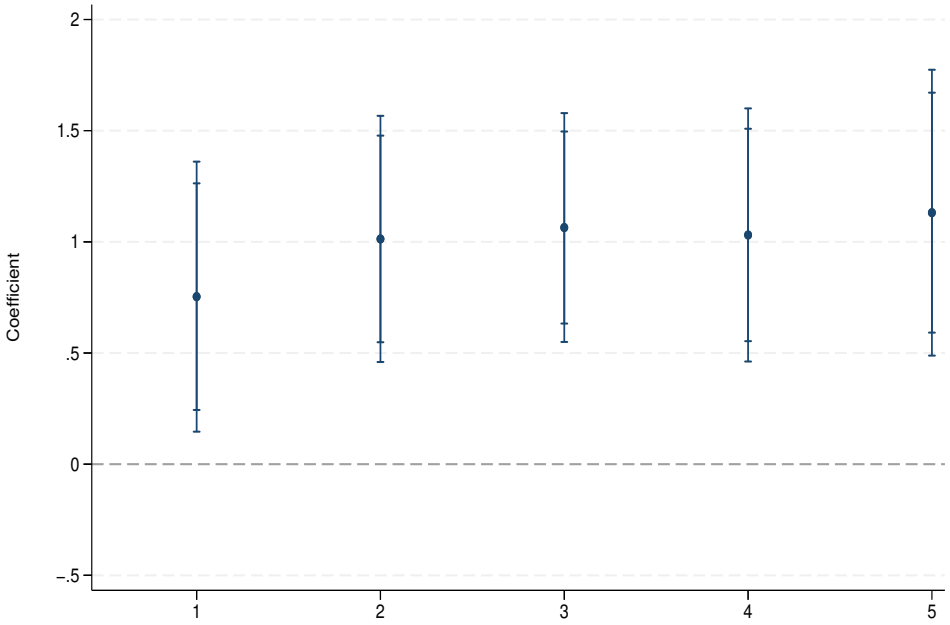
Notes. This Figure overlays the event-study plot using a dynamic version of the TWFE model. The outcome variable is the exit rate from social media. Exit is defined as a transition from active engagement to disconnection, where a user either moves from active (status = 1) to inactive. The exit rate is calculated as the average of platform-specific exit indicators (Instagram, Facebook, Twitter, and YouTube). For a detailed description of the outcome, treatment, and control variables, see Table A16. The bars represent 95% and 90% confidence intervals. Standard errors are clustered at the individual level.

Figure A12: Effects of the Instagram's algorithm on the amount of time spent on the Internet across different devices



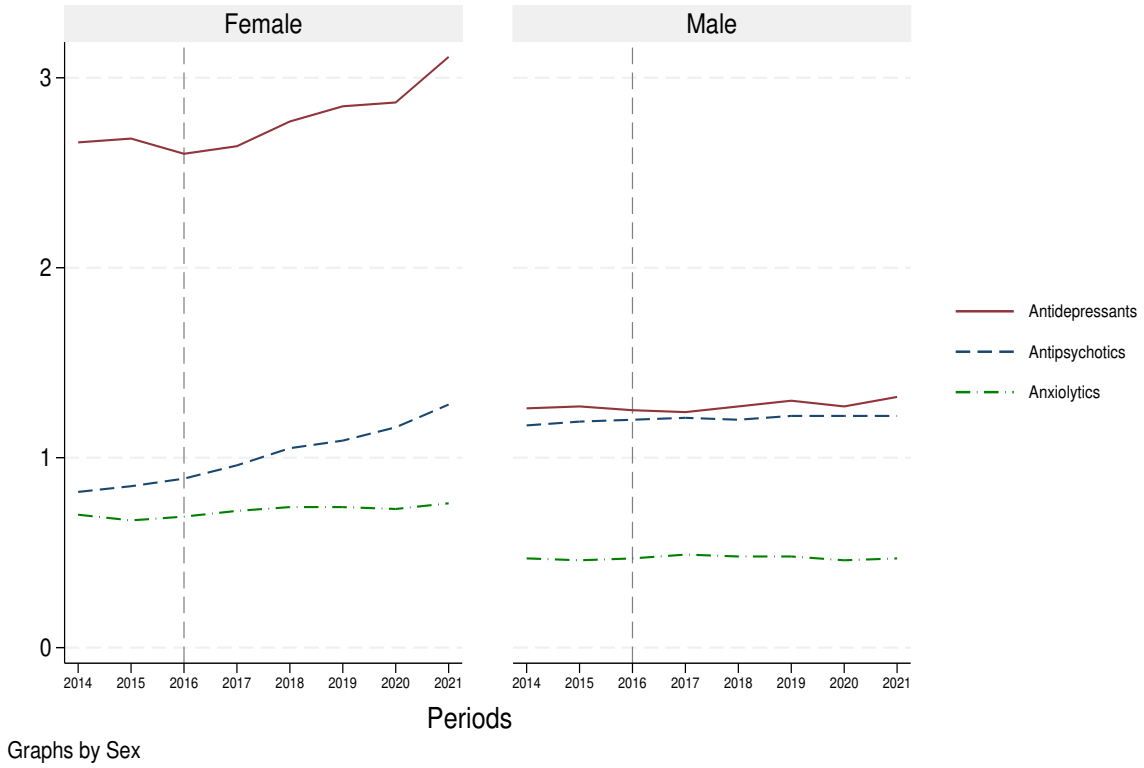
Notes. This Figure explores the effects of the introduction of the Instagram's algorithm on average hours spent on the Internet through different devices and an index of hours spent online across different devices. All outcomes are standardized so that, in the preperiod, they have a mean of zero and a standard deviation of one. All estimates are obtained using my preferred specification, namely the one including controls and individual and time fixed effects. My controls consist of net income, level of education, housing situation and level of urbanization. For a detailed description of the outcome, treatment, and control variables, see Table A16. The bars represent 95 and 90 percent confidence intervals. Standard errors are clustered at the individual level.

Figure A13: Effect of the Instagram's algorithm on the Index of Poor Mental Health by quantiles of hours spent on social media



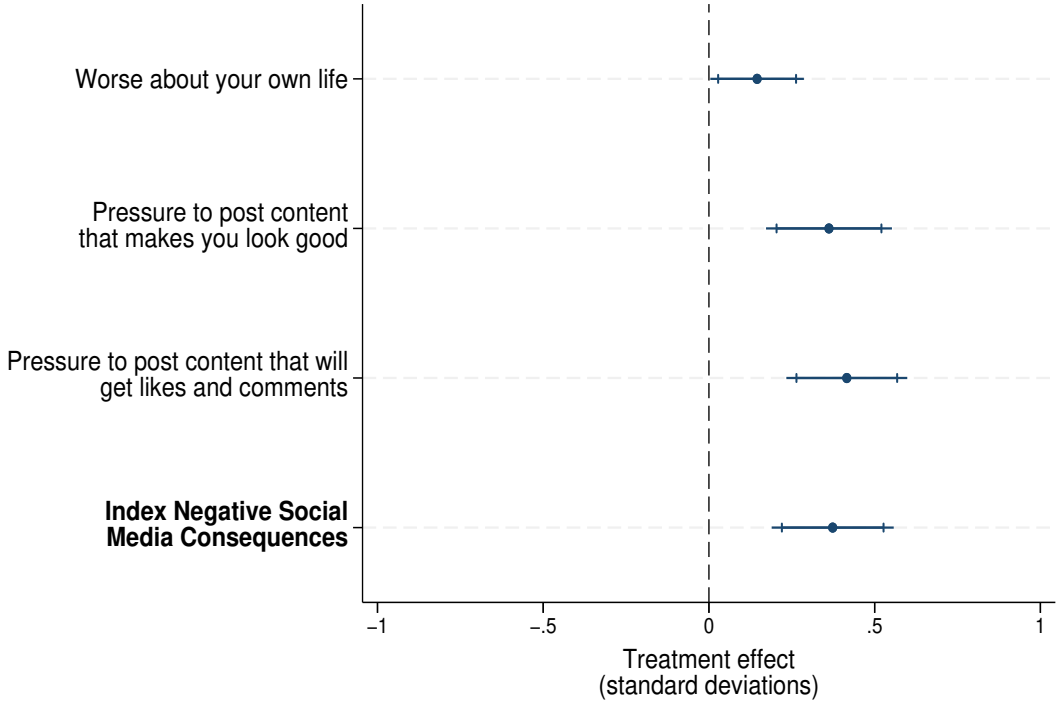
Notes. This Figure explores the effects of time spent on social media on my index of poor mental health. The index is standardized so that, over the previous period, it has a mean of zero and a standard deviation of one. The estimates are obtained using my preferred specification, namely the one including controls and individual and time fixed effects. My controls consist of net income, level of education, housing situation and level of urbanization. For a detailed description of the outcome, treatment, and control variables, see Table A16. The bars represent 95 and 90 percent confidence intervals. Standard errors are clustered at the individual level.

Figure A14: Percentage of people aged 15-25 years who were prescribed antidepressant, anxiolytic and antipsychotic drugs in the Netherlands, 2014-2021



Notes. This Figure shows the percentage of individuals aged 15-25 years to whom in the year concerned medicines were dispensed for which the costs are reimbursed under the statutory basic medical insurance. The data come from Statistics Netherlands (Centraal Bureau voor de Statistiek). The population includes everybody registered in the Personal Records Database (BRP) and living in the Netherlands at some point in the year concerned. For a detailed description of the outcome, treatment, and control variables, see Table A16.

Figure A15: External validity - evidence from Pew Research Center data



Notes. This Figure explores the effects of the introduction of the Instagram's algorithm on variables measuring the negative effects of social media using data from the Pew Research Center. All outcomes are standardized so that, in the preperiod, they have a mean of zero and a standard deviation of one. This specification include year of birth, age-group and region fixed effects. All estimates are obtained using my preferred specification, namely the one including controls and individual and time fixed effects. My controls consist of indicators for race (white, Hispanic, other non-Hispanic, and black), indicator for income (being above the median), region and gender. For a detailed description of the outcome, treatment, and control variables, see Table A16. The bars represent 95 and 90 percent confidence intervals. Standard errors are clustered at the region-age-wave level.

A.2 Appendix Tables

Table A1: Summary statistics by Instagram

	(1) No IG mean	(2) IG mean
Panel A. Baseline Characteristics		
Female	0.53	0.67
Dutch	0.81	0.72
Year of birth	1978	1985
Level of urbanization	2.88	2.89
Net income	1143.55	689.98
Housing situation	1.29	1.30
Level of education	4.02	3.49
Panel B. Baseline Mental Health		
Index Poor Mental Health	0.07	0.23
Index Symptoms Poor Mental Health	0.10	0.37
Index Depression Services	-0.01	-0.06
Observations	2749	1539

Notes. This table presents individual level summary statistics by treatment and control group. The data consists of individual-level characteristics retrieved from the LISS dataset. All indices are standardized so that, in the pre-period, they have a mean of zero and a standard deviation of one. For a detailed description of the outcome, treatment, and control variables, see Table A16.

Table A2: Summary statistics by Instagram (teenagers)

	(1) No IG	(2) IG
	mean	mean
Panel A. Baseline Characteristics		
Female	0.38	0.69
Dutch	0.51	0.49
Year of birth	1996	1996
Level of urbanization	3.24	3.18
Net income	7.98	32.10
Housing situation	1.22	1.14
Level of education	1.86	1.87
Panel B. Baseline Mental Health		
Index Poor Mental Health	0.28	0.62
Index Symptoms Poor Mental Health	0.57	0.90
Index Depression Services	-0.26	-0.04
Observations	86	313

Notes. This table presents individual level summary statistics by treatment and control group only for teenagers (individuals who are born between 1995 and 2000). The data consists of individual-level characteristics retrieved from the LISS dataset. All indices are standardized so that, in the pre-period, they have a mean of zero and a standard deviation of one. For a detailed description of the outcome, treatment, and control variables, see Table A16.

Table A3: Effect of the Instagram's algorithm on practicing physical activities

	Dummy Sport	Hours	Index Sport
	(1)	(2)	(3)
Post IG Algorithm Introduction	-0.020	-0.436	-0.065
	(0.094)	(0.909)	(0.170)
Baseline mean	0.69	5.57	-0.23
Observations	1,043	687	1,043
Individual fixed effects	✓	✓	✓
Time fixed effects	✓	✓	✓
Controls	✓	✓	✓

Notes. This table explores the effects of the introduction of the Instagram algorithm on practicing physical activities. It display estimates of coefficient β from equation 1 using my preferred specification, namely the one including controls and individual and time fixed effects. All indices are standardized so that, in the pre-period, they have a mean of zero and a standard deviation of one. My controls consist of net income, level of education, housing situation and level of urbanization. For a detailed description of the outcome, treatment, and control variables, see Table A16. Standard errors in parentheses are clustered at the individual level. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table A4: Effect of the Instagram's algorithm on the Index of Poor Mental Health - without Tweeter users

	Index of Poor Mental Health	
	(1)	(2)
Cohort 1995-2000:		
Post IG Algorithm Introduction	0.441**	0.509**
	(0.187)	(0.231)
Cohort 1985-1994:		
Post IG Algorithm Introduction	-0.008	-0.002
	(0.079)	(0.089)
Cohort 1976-1984		
Post IG Algorithm Introduction	-0.034	-0.056
	(0.071)	(0.075)
Individual fixed effects	✓	✓
Time fixed effects	✓	✓
Controls		✓
Observations	5,847	5,412

Notes. This table explores the effect of the introduction of the algorithm on Instagram on individuals' mental health. It presents estimates of coefficient β from equation 1 with my index of poor mental health as outcome variable. Specifically, I exclude from the sample those individuals having a tweeter account. The index is standardized so that, in the preperiod, it has a mean of zero and a standard deviation of one. Column 1 estimates equation 1 without including controls column 2 estimates equation 1 including controls. My controls consist of net income, level of education, housing situation and level of urbanization. For a detailed description of the outcome, treatment, and control variables, see Table A16. Standard errors in parentheses are clustered at the individual level. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table A5: Robustness with other social media platforms

	Index of Poor Mental Health	
	(1)	(2)
Treatment: FB account		
Post IG Algorithm Introduction	-0.190	-0.133
	(0.180)	(0.217)
Treatment: TW/X account		
Post IG Algorithm Introduction	-0.087	-0.368
	(0.352)	(0.383)
Treatment: YT account		
Post IG Algorithm Introduction	-0.001	-0.082
	(0.677)	(0.633)
Individual fixed effects	✓	✓
Time fixed effects	✓	✓
Controls		✓

Notes. This table displays estimates of coefficient β from equation 1, using different groups of treated and control units, as falsification tests. Specifically, in the first row, I estimate equation 1 using as treatment having a Facebook account versus having a social media account other than Facebook. In the second row, I estimate equation 1 using as treatment having a Twitter/X account versus having a social media account other than Twitter/X. In the third row, I estimate equation 1 using as treatment having a YouTube account versus having a social media account other than YouTube, always excluding those with Instagram accounts. The outcome variable is my index of poor mental health. The index is standardized so that, in the preperiod, it has a mean of zero and a standard deviation of one. Column 1 estimates equation 1 without including controls column 2 estimates equation 1 including controls. My controls consist of net income, level of education, housing situation and level of urbanization. For a detailed description of the outcome, treatment, and control variables, see Table A16. Standard errors in parentheses are clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A6: Effect of the Instagram's algorithm on the Index of Poor Physical Health

	Index of Poor Physical Health	
	(1)	(2)
Post IG Algorithm Introduction	-0.145	-0.177
	(0.247)	(0.326)
Observations	1,080	1,001
Individual fixed effects	✓	✓
Time fixed effects	✓	✓
Controls		✓

Notes. This table explores the effect of the introduction of the algorithm on Instagram on individuals' physical health. It presents estimates of coefficient β from equation 1 with my index of poor physical health as outcome variable. Physical health is an index constructed as follows: I orient all variables so that higher values indicate worse physical health outcomes. Then, I standardize these variables using means and standard deviations from the preperiod. Next, I calculate an equally weighted average of the index components and finally I standardize the final index. Column 1 estimates equation 1 without including controls column 2 estimates equation 1 including controls. My controls consist of net income, level of education, housing situation and level of urbanization. For a detailed description of the outcome, treatment, and control variables, see Table A16. Standard errors in parentheses are clustered at the individual level. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table A7: Alternative index construction methods

	Include obs. with missing values	Equally-weighted index	Anderson (2008)
	(1)	(2)	(3)
Post IG Algorithm Introduction	0.394**	0.346*	0.444**
	(0.181)	(0.193)	(0.175)
Observations	1,045	1,028	1,045
Individual fixed effects	✓	✓	✓
Time fixed effects	✓	✓	✓
Controls	✓	✓	✓

Notes. This table presents estimates of coefficient β from equation 1, using my preferred specification, namely the one including controls and individual and time fixed effects. Specifically, it shows robustness checks to different ways of constructing the index of poor mental health. Column (1) presents my baseline results, which rely on the index construction method described in Section 3. Column (2) presents results on a version of the index that does not include observations for which some of the index components are missing. Column (3) presents results on an inverse-covariance weighted index that assigns a smaller weight to strongly correlated components (Anderson (2008)). My controls consist of net income, level of education, housing situation and level of urbanization. For a detailed description of the outcome, treatment, and control variables, see Table A16. Standard errors in parentheses are clustered at the individual level. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table A8: Balance

Variable	(1) No IG Mean/(SE)	(2) Yes IG Mean/(SE)	(1)-(2) Pairwise t-test Mean difference
Negative financial expectations for the coming year	3.25	2.80	0.45
	(0.48)	(0.20)	
Confidence in reaching age 75	7.85	7.73	0.12
	(0.17)	(0.10)	
Confidence in reaching age 80	7.14	6.93	0.21
	(0.20)	(0.11)	
Satisfaction with the way they look	5.43	4.69	0.75
	(0.13)	(0.10)	
I have equal worth to others	5.49	5.30	0.19
	(0.17)	(0.12)	
Chance of finding a job in the coming year	766.75	627.40	139.35
	(232.25)	(227.65)	
Am the life of the party	3.46	3.61	-0.15
	(0.12)	(0.06)	
Don't talk a lot	2.83	2.32	0.51
	(0.15)	(0.08)	
Feel comfortable around people	3.94	4.07	-0.12
	(0.10)	(0.05)	
Keep in the background	3.26	2.78	0.48
	(0.14)	(0.08)	
Index Poor Mental Health	0.28	0.62	-0.35
	(0.07)	(0.06)	

Notes. This table presents a balance table on the following variables: financial expectations for the future, confidence in reaching ages 75 and 80, satisfaction with appearance, sense of equal worth to others, perceived job prospects in the coming year, self-perception as the "life of the party," tendency to be reserved, comfort around others, inclination to remain in the background, and my general index of poor mental health. The first column shows the mean value of the demographic characteristics in the pre-period for the control group; the second columns shows the mean value of those characteristics in the pre-period for the treatment group. The p-values are calculated after residualizing each characteristic on individual and time fixed effects. For a detailed description of the outcome, treatment, and control variables, see Table A16. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A9: Effect of the Instagram's algorithm on the Index of Poor Mental Health - with treatment defined as Instagram users until October 2015

	Index of Poor Mental Health	
	(1)	(2)
Post IG Algorithm Introduction	0.258*	0.341*
	(0.151)	(0.190)
Observations	773	712
Individual fixed effects	✓	✓
Time fixed effects	✓	✓
Controls		✓

Notes. This table explores the effect of the introduction of the algorithm on Instagram on individuals' mental health where the treatment is constructed as individuals who reported having an Instagram account by October 2015. It presents estimates of coefficient β from equation 1 with my index of poor mental health as outcome variable. The index is standardized so that, in the preperiod, it has a mean of zero and a standard deviation of one. Column 1 estimates equation 1 without including controls, while column 2 includes controls. My controls consist of net income, level of education, housing situation and level of urbanization. For a detailed description of the outcome, treatment, and control variables, see Table A16. Standard errors in parentheses are clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A10: Baseline Characteristics by Instagram Adoption Timing (Before 2016 vs. 2016)

Variable	(1) IG Before 2016 Mean/(SE)	(2) IG 2016 Mean/(SE)	(1)-(2) Pairwise t-test Mean difference
Female	0.69	0.66	.
	(0.03)	(0.07)	
Dutch	0.49	0.51	.
	(0.03)	(0.07)	
Year of birth	1,996.43	1,996.09	.
	(0.07)	(0.16)	
Level of urbanization	3.15	3.37	-0.22**
	(0.08)	(0.17)	
Net income	32.41	30.34	2.07
	(6.11)	(16.35)	
Housing situation	1.13	1.23	-0.11
	(0.02)	(0.08)	
Level of education	1.84	2.02	-0.18**
	(0.06)	(0.14)	
Index Poor Mental Health	0.63	0.59	0.04
	(0.07)	(0.14)	
Index Symptoms Poor Mental Health	0.90	0.91	-0.01
	(0.07)	(0.16)	
Index Depression Services	-0.03	-0.10	0.08
	(0.06)	(0.11)	

Notes. This table presents a balance table on the following variables: gender (female), nationality (Dutch), age, level of urbanization, net income, housing situation, level of education, general index of poor mental health, index of symptoms of poor mental health, index of poor physical health, and index of depression services. The table compares individuals who reported having an Instagram account by 2015 with those who reported joining Instagram in 2016 (treatment group). The first column shows the mean value of these characteristics in the pre-period for the former group; the second column shows the mean value of these characteristics in the pre-period for the latter group. The p-values are calculated after residualizing each characteristic on individual and time fixed effects. For a detailed description of the outcome, treatment, and control variables, see Table A16. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A11: Effect of the Instagram's algorithm on the Index of Poor Mental Health - excluding the period of the Covid-19 pandemic

	Index of Poor Mental Health	
	(1)	(2)
Post IG Algorithm Introduction	0.269*	0.350**
	(0.140)	(0.176)
Observations	955	888
Individual fixed effects	✓	✓
Time fixed effects	✓	✓
Controls		✓

Notes. This table explores the effect of the introduction of the algorithm on Instagram on individuals' mental health. It presents estimates of coefficient β from equation 1 with my index of poor mental health as outcome variable. Specifically, I exclude 2020 and 2021, to check whether my results hold when excluding the years of the Covid-19 pandemic. The index is standardized so that, in the preperiod, it has a mean of zero and a standard deviation of one. Column 1 estimates equation 1 without including controls column 2 estimates equation 1 including controls. My controls consist of net income, level of education, housing situation and level of urbanization. For a detailed description of the outcome, treatment, and control variables, see Table A16. Standard errors in parentheses are clustered at the individual level. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table A12: Effect of the Instagram's algorithm on missing values

	Any Missing Values	Total Missing Values	Index of Missing Values
	(1)	(2)	(3)
Post IG Algorithm Introduction	-0.010	-0.083	-0.272
	(0.046)	(0.157)	(0.515)
Baseline mean	0.02	0.04	0.00
Observations	1,045	1,045	1,045
Individual fixed effects	✓	✓	✓
Time fixed effects	✓	✓	✓
Controls	✓	✓	✓

Notes. This table addresses the potential reduction in the stigma associated with mental issues as a result of the introduction of the algorithm on Instagram. Specifically, it presents estimates of coefficient β from equation 1, using my preferred specification, namely the one including controls and individual and time fixed effects, with three different ways of aggregating missing responses. In Column (1), the outcome is an indicator equal to one if a respondent did not answer at least one question composing the index of poor mental health, and equal to zero otherwise. In Column (2), the outcome is the total number of questions composing the index of poor mental health left unanswered by a respondent. In Column (3) the number of unanswered questions is standardized using means and standard deviations from the pre-period. My controls consist of net income, level of education, housing situation and level of urbanization. For a detailed description of the outcome, treatment, and control variables, see Table A16. Standard errors in parentheses are clustered at the individual level. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table A13: Effects of the Instagram's algorithm on missing values for each mental health questions

	Taking medicine for anxiety or depression	Taking medicine for sleeping problems	Therapy depression	Eating disorder	Last year felt anxious	Last year felt very sad	Last year felt depressed and gloomy
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Post IG Algorithm Introduction	0.019	0.019	-0.036	0.009	-0.031	-0.031	-0.031
	(0.024)	(0.024)	(0.038)	(0.008)	(0.037)	(0.037)	(0.037)
Baseline mean	0.01	0.01	0.01	0.00	0.00	0.00	0.00
Observations	1,045	1,045	1,045	1,045	1,045	1,045	1,045
Individual fixed effects	✓	✓	✓	✓	✓	✓	✓
Time fixed effects	✓	✓	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓	✓	✓

Notes. This table addresses the potential reduction in the stigma associated with mental issues as a result of the introduction of the algorithm on Instagram. Specifically, it presents estimates of coefficient β from equation 1 using my preferred specification, namely the one including controls and individual and time fixed effects. The outcome variables are the number of unanswered questions to each mental health questions, which are standardized using means and standard deviations from the pre-period. My controls consist of net income, level of education, housing situation and level of urbanization. For a detailed description of the outcome, treatment, and control variables, see Table A16. Standard errors in parentheses are clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A14: Effects of the Instagram's algorithm on smoking and substance use

	Cigarettes	Pipe	Cigars	E-cigarettes	Sedatives	Soft drugs	XTC	Hallucinogens	Hard drugs	Index
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Post IG Algorithm Introduction	-0.137	0.000	-0.000	-0.009	0.087	-0.122	-0.040	0.001	-0.030	-0.193
	(0.092)	(.)	(0.008)	(0.012)	(0.051)	(0.075)	(0.046)	(0.010)	(0.035)	(0.260)
Baseline mean	0.22	0.00	0.00	0.01	0.02	0.08	0.01	0.00	0.01	0.00
Observations	1,045	1,045	1,045	1,045	1,041	1,041	1,041	1,041	1,041	1,045
Individual fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Time fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Notes. This table explores the effects of the introduction of the Instagram algorithm on individuals' self-reported behaviors related to smoking and substance use. It display estimates of coefficient β from equation 1 using my preferred specification, namely the one including controls and individual and time fixed effects. All variables are standardized so that, in the pre-period, they have a mean of zero and a standard deviation of one. My controls consist of net income, level of education, housing situation and level of urbanization. For a detailed description of the outcome, treatment, and control variables, see Table A16. Standard errors in parentheses are clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A15: Effect of the Instagram's algorithm on alcohol consumption

	Drink count	Used 30 days	Used daily	Index std. dev.
	(1)	(2)	(3)	(4)
Post IG Algorithm Introduction	-1.439	-0.075	-0.112	-0.665
	(1.165)	(0.100)	(0.055)	(0.303)
Baseline mean	3.19	0.67	0.02	-0.00
Observations	1,045	1,041	1,041	1,045
Individual fixed effects	✓	✓	✓	✓
Time fixed effects	✓	✓	✓	✓
Controls	✓	✓	✓	✓

Notes. This table explores the effects of the introduction of the Instagram algorithm on individuals' self-reported behaviors related to alcohol. It display estimates of coefficient β from equation 1 using my preferred specification, namely the one including controls and individual and time fixed effects. All indices are standardized so that, in the pre-period, they have a mean of zero and a standard deviation of one. My controls consist of net income, level of education, housing situation and level of urbanization. For a detailed description of the outcome, treatment, and control variables, see Table A16. Standard errors in parentheses are clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A16: Variables definitions, constructions, and associated LISS survey questions

Variable	Description
Treatment Variables	
Post	Coding: 1 = The survey was conducted after the introduction of the algorithm on Instagram (March 2016); 0 = The survey was conducted before the introduction of the algorithm.
Instagram	Question:"Which of the following social media do you use? Instagram" Coding: 1 = Individuals claim to use it by 2016; 0 = Individuals claim to have at least one social media account by 2016 on other social media platforms such as Facebook, Youtube and/or Twitter.
Facebook	Question:"Which of the following social media do you use? Facebook" Coding: 1 = Individuals claim to use it by 2016; 0 = otherwise.
Twitter	Question:"Which of the following social media do you use? Twitter" Coding: 1 = Individuals claim to use it by 2016; 0 = otherwise.
Youtube	Question:"Which of the following social media do you use? Youtube" Coding: 1 = Individuals claim to use it by 2016; 0 = otherwise.
Main Indices	
Index Poor Mental Health	The index is constructed as follows: (1) I standardized all variables related to symptoms of poor mental health (see below) and all variables related to depression services (see below) so that they have a mean of 0 and a standard deviation of 1 in the pre-period; (2) I took an equally weighted average of the standardized variables; (3) I re-standardized the equally weighted average so that it has a mean of 0 and a standard deviation of 1 in the pre-period.
Index Symptoms Poor Mental Health	Similar construction as above but focusing only on symptoms of poor mental health.
Index Depression Services	The index is constructed by standardizing variables related to depression services and averaging them as above.
Index Symptoms Depression	Focused on variables related to symptoms of depression; standardized and averaged as above.
Index Symptoms Other Conditions	Focused on variables related to symptoms of other mental health conditions; standardized and averaged similarly.
Index Symptoms Poor Mental Health	
Index Symptoms Depression	
Past mont felt anxious	Question: "Past month I felt very anxious"; Scale: 1 = never, 2 = seldom, 3 = sometimes, 4 = often, 5 = mostly, 6 = continuously.

Variable	Description
Past mont felt very sad	Question: "Past month I felt so down that nothing could cheer me up"; Scale: same as above.
Past mont felt depressed and gloomy	Question: "Past month I felt depressed and gloomy"; Scale: same as above.
Index of Other Conditions	
Eating disorder	The variable is constructed by calculating the BMI from height and weight, and then classified as follows: dca = 4 if BMI $<$ 16 (indicative of anorexia), dca = 3 if 16 $\leq$ BMI $<$ 18.5, dca = 2 if 18.5 $\leq$ BMI $<$ 25, dca = 1 if BMI $>$ 25, and dca = . (missing) if dca == 0.
Depression Services	
Taking medicines for anxiety or depression	Question: "Are you currently taking medicine at least once a week for anxiety or depression?"; Scale: 1 = yes; 0 = no.
Taking medicines for sleeping problems	Question: "Are you currently taking medicine at least once a week for sleeping problems?"; Scale: same as above.
Therapy depression	Question: "With what specialist(s) did you have contact over the past 12 months? psychiatrist"; Scale: same as above.
Controls and socio-demographic characteristics	
Net income	Personal net monthly income in categories; Scale: 0 = No income, 1 = EUR 500 or less, 2 = EUR 501 to EUR 1000, 3 = EUR 1001 to EUR 1500, 4 = EUR 1501 to EUR 2000, 5 = EUR 2001 to EUR 2500, 6 = EUR 2501 to EUR 3000, 7 = EUR 3001 to EUR 3500, 8 = EUR 3501 to EUR 4000, 9 = EUR 4001 to EUR 4500, 10 = EUR 4501 to EUR 5000, 11 = EUR 5001 to EUR 7500, 12 = More than EUR 7500
Level of education	Highest level of education irrespective of diploma; Scale: 1 = primary school, 2 = vmbo (intermediate secondary education), 3 = havo/vwo (higher secondary education), 4 = mbo (intermediate vocational education), 5 = hbo (higher vocational education), 6 = wo (university), 7 = other, 9 = not (yet) started any education
Housing situation	Housing situation; Scale; 1 = self-owned dwelling, 2 = rental dwelling, 3 = cost-free dwelling
Female	Gender; Scale: 1 = female, 0 = male
Dutch	Origin; Scale: 1 = dutch, 0 = otherwise
Year of birth	Year of birth
Level of urbanization	Level of urbanization; Scale: 1 = extremely urban, 2 = very urban, 3 = moderately urban, 4 = slightly urban, 5 = not urban.

Variable	Description
Downstream Effects	
Daily activities	Question: "To what extent did your physical health or emotional problems hinder your daily activities over the past month, for instance in going for a walk, walking up stairs, dressing yourself, washing yourself, visiting the toilet?"; Scale: 1 = not at all, 2 = hardly, 3 = a bit, 4 = quite a lot, 5 = very much.
Social activities	Question: "To what extent did your physical health or emotional problems hinder your social activities over the past month, such as visiting friends and acquaintances?"; Scale: same as above.
Work/Study	Question: "To what extent did your physical health or emotional problems hinder your work over the past month, for instance in your job, the housekeeping, or in school?"; Scale: same as above.
Index Downstream Effects	The index is constructed aggregating the variables above following the same procedure as the Index Poor Mental Health.
Social Comparison	
Social validation	Question:"Which values act as a guiding principle in your life and which values are less important to you? social recognition"; Scale: 1 = extremely unimportant, 7 = extremely important.
Low self-esteem	Question:"At times, I think I am no good at all"; Scale: 1 = totally disagree, 7 = totally agree.
Low self-worth	Question:"I feel that I'm a person of worth, at least on an equal plan with other"; Scale: same as above.
Index Self-Perception	The index is constructed aggregating the variables above following the same procedure as the Index Poor Mental Health.
Index Social Comparison	
Jealousy of the good fortune of others	Question:"There have been times when I was quite jealous of the good fortune of others"; Scale: 1 = True , 0 = False. The average value for the pre-period is then calculated, and a dummy variable is created to indicate whether this value is above the median. The final scale is: 1 = above the median, 0 = below the median.
Avoid being the center of attention	Question:"Don't mind being the center of attention"; Scale: 1 = very inaccurate, 2 = moderately inaccurate, 3 = neither inaccurate nor accurate, 4 = moderately accurate, 5 = very accurate. The variable is adjusted so that higher values indicate a negative outcome. The average value for the pre-period is then calculated, and a dummy variable is created to indicate whether this value is above the median. The final scale is: 1 = above the median, 0 = below the median.

Variable	Description
Do not take time out for others	Question:"Take time out for others"; Scale: 1 = very inaccurate, 2 = moderately inaccurate, 3 = neither inaccurate nor accurate, 4 = moderately accurate, 5 = very accurate. The variable is adjusted so that higher values indicate a negative outcome. The average value for the pre-period is then calculated, and a dummy variable is created to indicate whether this value is above the median. The final scale is: 1 = above the median, 0 = below the median.
Not relaxed most of the time	Question:"Am relaxed most of the time"; Scale: 1 = very inaccurate, 2 = moderately inaccurate, 3 = neither inaccurate nor accurate, 4 = moderately accurate, 5 = very accurate. The variable is adjusted so that higher values indicate a negative outcome. The average value for the pre-period is then calculated, and a dummy variable is created to indicate whether this value is above the median. The final scale is: 1 = above the median, 0 = below the median.
Index Social Comparison	Coding: Index sums the binary variables defined above. As an additional moderator to study heterogeneous treatment effects, I consider whether an individual is above the median value of the index of social comparisons or below the median value.
Social media trap	
People are less inclined to meet in real life	Question:" Because of social media people are less inclined to meet in real life"; Scale: 1 = completely disagree, 5 = completely agree.
It is easier for people to maintain friendships	Question:"Because of social media sites it is easier for people to maintain friendships"; Scale: same as above.
Using social media is harmful for the real social life	Question:"When people do not use social media, this is harmful for the 'real' (non-virtual) social life. Scale: same as above. The variable is adjusted so that higher values indicate a negative outcome.
Index Social Media Consequences	The index is constructed aggregating the variables above following the same procedure as the Index Poor Mental Health.
Exit rate	The exit rate measures the proportion of social media platforms a user stops using relative to the total number of platforms they were using in the previous wave. A user is considered to have stopped using a platform when their status transitions from active to inactive. Indicators are created for Instagram, Facebook, Twitter, and YouTube to capture this behavior. The total number of platforms a user stops using is calculated, along with the total number of platforms they were previously active on. The exit rate is then computed as the ratio of platforms stopped to platforms previously active.
Disruptive Internet Use	
Time spent online on the following activies	

Variable	Description
Tablet	Question:"Can you indicate how many hours you use the Internet on a tablet per week, on average (including emailing), for other things than completing the questionnaires of this panel?"
Smartphones	Question:"Can you indicate how many hours you use the Internet on a smartphone per week, on average (including emailing), for other things than completing the questionnaires of this panel?"
Index Hours Internet	The index is constructed aggregating the variables above following the same procedure as the Index Poor Mental Health.
Time spent online across devices	
E-mail	Question:"Can you indicate how many hours per week, on average, you spend on these online activities?"; Scale: 0..168 hours per week.
Searching for information on the Internet (e.g. about hobbies, work, opening hours, daytrips, etc.)	Question:"Can you indicate how many hours per week, on average, you spend on these online activities?; Scale: same as above.
Searching for and comparing products/product information on the Internet	Question:"Can you indicate how many hours per week, on average, you spend on these online activities?"; Scale: same as above.
Purchasing items via the Internet	Question:"Can you indicate how many hours per week, on average, you spend on these online activities?"; Scale: same as above.
Watching online films or TV programs	Question:"Can you indicate how many hours per week, on average, you spend on these online activities?"; Scale: same as above.
Downloading software, music or films	Question:"Can you indicate how many hours per week, on average, you spend on these online activities?"; Scale: same as above.
Internet banking	Question:"Can you indicate how many hours per week, on average, you spend on these online activities?"
Playing Internet games/online gaming	Question:"Can you indicate how many hours per week, on average, you spend on these online activities?"; Scale: same as above.
Reading online news and magazines	Question:"Can you indicate how many hours per week, on average, you spend on these online activities?"; Scale: same as above.
Newsgroups	Question:"Can you indicate how many hours per week, on average, you spend on these online activities?"; Scale: same as above.

Variable	Description
Reading and viewing social media (e.g., Facebook, Instagram, Twitter, YouTube, LinkedIn, Google+, Pinterest, Flickr, or similar services)	Question:"Can you indicate how many hours per week, on average, you spend on these online activities?"; Scale: same as above.
Reading and/or writing blogs	Question:"Can you indicate how many hours per week, on average, you spend on these online activities?"; Scale: same as above.
Posting messages, photos and short films on social media yourself (e.g., Facebook, Instagram, Twitter, YouTube, LinkedIn, Google+, Pinterest, Flickr, or similar services)	Question:"Can you indicate how many hours per week, on average, you spend on these online activities?"; Scale: same as above.
Chatting, video calling or sending messages via WhatsApp, Telegram, Snapchat, Skype or similar services	Question:"Can you indicate how many hours per week, on average, you spend on these online activities?"; Scale: same as above.
Dating websites (like Relatieplanet, Lexa, Tinder, Grindr or similar services)	Question:"Can you indicate how many hours per week, on average, you spend on these online activities?"; Scale: same as above.
Visiting (discussion) forums and Internet communities	
Other activities on the Internet	Question:"Can you indicate how many hours per week, on average, you spend on these online activities?"; Scale: same as above.
Index Online Activities	The index is constructed aggregating the variables above following the same procedure as the Index Poor Mental Health.
Missing Values Variables	
Any missing values	1 = respondent left unanswered at least one question composing the index of poor mental health; 0 = respondent answered all the questions composing the index of poor mental health.

Variable	Description
Total missing values	The number of questions composing the index of poor mental health that a respondent left unanswered. Index of missing values The index is constructed as follows: (1) I consider all variables that comprise the index of poor mental health (2) I calculate the total number of question that respondent left unanswered (3) I standardized the total so it has a mean of 0 and standard deviation of 1 in the pre-period.
Index of missing values	The index is constructed as follows: (1) I consider all variables that comprise the index of poor mental health (2) I calculate the total number of question that respondent left unanswered (2) I standardized the total so it has a mean of 0 and standard deviation of 1 in the pre-period.
Missing values	I create a new variables for each one composing the index of poor mental health: 1 = respondent left unanswered the question; 0 = respondent answered.
Drinking	
Drink count	It considers different questions:
Beer	Question:"Can you indicate below how much beer (of normal strength, pilsner, white beer, dark beer, containing less than 6% alcohol) you drank the one day during the last week on which you drank the most amount of drinks containing alcohol?"; Answers: (1) number of glasses (count large glasses as 2); (2)number of half liter glasses (pints); (3) number of half liter cans or bottles; (4) number of small cans or bottles; Scale: 0..9999, empty.
Strong beer	Question:"Can you indicate below how much strong beer (special beers with 6% alcohol or more),you drank the one day during the last week on which you drank the most amount of drinks containing alcohol?"; Answers: (1) number of glasses (count large glasses as 2); (2)number of half liter glasses (pints); (3) number of half liter cans or bottles; (4) number of small cans or bottles; Scale: same as above.
Alcoholic beverages	Question: "Can you indicate below how many of these alcoholic beverages you drank the one day during the last week on which you drank the most amount of drinks containing alcohol?" ; Answers: (1) strong spirits or liquor, such as gin, whisky, rum, brandy, vodka or cocktails; (2) sherry or martini (including port, vermouth, Cinzano, Dubonnet); (3) wine (including champagne); Scale: 1..9999.
Premixes, alcohol pops, blasters and shooters	Question:"Can you indicate below how many small cans or bottles of premixes, alcohol pops, blasters and shooters (such as Bacardi Breezer, Smirnoff Ice) you drank the one day during the last week on which you drank the most amount of drinks containing alcohol?" ; Scale: 1..9999 small cans or bottles
Other type of alcoholic drink (1)	Question:"Can you indicate below how many glasses you drank the most amount of drinks containing alcohol?" ; Scale: 1..9999 glasses

Variable	Description
Other type of alcoholic drink (2)	Question:"Can you indicate below how many glasses you drank the most amount of drinks containing alcohol?" ; Scale: 1..9999 glasses
Used 30 days	Question:"Now think of all the sorts of drink that exist. How often did you have a drink containing alcohol over the last 12 months?"; Answers: (1) almost every day, (2) five or six days per week, (3) three or four days per week; (4) once or twice a week, (5) once or twice a month, (6) once every two months, (7) once or twice a year, (8) not at all over the last 12 months; Scale: 1= if ≤ 5, 0= if > 5.
Used daily	Question:"Now think of all the sorts of drink that exist. How often did you have a drink containing alcohol over the last 12 months?"; Answers: (1) almost every day, (2) five or six days per week, (3) three or four days per week; (4) once or twice a week, (5) once or twice a month, (6) once every two months, (7) once or twice a year, (8) not at all over the last 12 months; Scale: 1= if ≤ 2, 0= if > 2.
Index	The index is constructed as follows: (1) I standardized the three variables above so that they have a mean of 0 and a standard deviation of 1 in the pre-period. (2) I took an equally-weighted average of the standardized variables. (3)) I re-standardized the equally-weighted average so that it has a mean of 0 and a standard deviation of 1 in the pre-period.
Smoking and Substance use	
Smoking	Question:"What do you smoke?"; Answers: (1) cigarettes (including rolling tobacco); (2) pipe; (3) cigars or cigarillos; (4) e-cigarettes (More than one answer possible); Scale: 1 = yes, 0 = no.
Substance use	Question:"Did you use one or more of the following substances over the past month?"; Answers: (1) sedatives (such as valium); (2) soft drugs (such as hashish, marijuana); (3) XTC; (4) hallucinogens (such as LSD, magic mushrooms); (5) hard drugs (such as stimulants, cocaine, heroin); Scale: 1 = never, 2 = sometimes, 3 = regularly.
Index	The index of substance use is constructed aggregating the variables above following the same procedure as the Index Poor Mental Health not discarding observations when one of the variables above is missing.
Practicing physical activities	
Dummy Sport	Question:"Do you practice sports?"; Scale: 1 = yes, 0 = no.
Hours	Question:"How many hours do you spend on sports per week, on average?"; Scale: 0.0..168.0 hours.

Variable	Description
Index Sport	The index is constructed aggregating the variables above following the same procedure as the Index Poor Mental Health.
Social activities	
Spend an evening with family	Question:"How often do you do the spend an evening with family (other than members of your own household)?"; Scale:1 = almost every day, 2 = once or twice a week, 3 = a few times per month, 4 = about once a month, 5 = a number of times per year, 6 = about once a year, 7 = never, 8 = don't know, 9 = not applicable.
Spend an evening with someone from the neighborhood	Question:"How often do you do the spend an evening with someone from the neighborhood?"; Scale: same as above.
Spend an evening with friends outside your neighborhood	Question:"How often do you do the spend an evening with friends outside your neighborhood?"; Scale: same as above.
Visit a bar or café	Question:"How often do you visit a bar or café?"; Scale: same as above.
Index Social Activities	The index is constructed aggregating the variables above following the same procedure as the Index Poor Mental Health.
Social connection quality measures	
Satisfaction with your social contacts	Question:"How satisfied are you with your social contacts?"; Scale: 0 = not at all satisfied, 10 = completely satisfied.
Enjoy your friends	Questions:"I enjoy my friends a lot"; Scale: 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, 5 = strongly agree.
Do not have a sense of emptiness around you	Question:"To what extent do the following statements apply to you, based on how you are feeling at present? I have a sense of emptiness around me"; Scale: 1 = no, 2 = more or less; 3 = yes.
There are enough people you can count on	Question:"To what extent do the following statements apply to you, based on how you are feeling at present? There are enough people I can count on in case of a misfortune"; Scale: 1 = yes, 2 = more or less; 3 = no.
You know a lot of people you can fully rely on	Question:"To what extent do the following statements apply to you, based on how you are feeling at present? I know a lot of people that I can fully rely on"; Scale: same as above.
There are enough people to whom you feel closely connected	Question:"To what extent do the following statements apply to you, based on how you are feeling at present? There are enough people to whom I feel closely connected"; Scale: same as above.

Variable	Description
You do not miss having people around you	Question:"To what extent do the following statements apply to you, based on how you are feeling at present? I miss having people around me"; Scale: 1 = no, 2 = more or less; 3 = yes.
You rarely feel deserted	Question:"To what extent do the following statements apply to you, based on how you are feeling at present? I often feel deserted"; Scale: same as above.
Index Social Connection	Index sums the binary variables defined above. As an additional moderator to study heterogeneous treatment effects, I consider whether a the median value of the index of social comparisons or below the median value.
Physical Health	
Index poor physical health	Physical health is an index composed of the following variables: general health perception (Scale: 1 = poor, 2 = moderate, 3 = good, 4 = very good, 5 = excellent); difficulties in mobility and daily activities (walking 100 meters, sitting for two hours, standing up from a chair, climbing stairs, crouching, or kneeling); difficulties with physical tasks (reaching above shoulder height, moving large objects, lifting or carrying 5 kilos, picking up a small coin, dressing or undressing, walking across a room, bathing or showering, eating, getting in and out of bed, using the toilet); difficulties with cognitive or instrumental activities (reading maps, preparing meals, shopping, using the phone, managing medicines, performing household tasks, maintaining the garden, managing finances) (Scale: 1 = without any trouble, 2 = with some trouble, 3 = with a lot of trouble, 4 = only with the help of others, 5 = not at all); and regular health complaints (joint pain, heart issues, breathing problems, flu-like symptoms, stomach issues, headaches, fatigue, sleeping difficulties, other recurrent complaints, or no recurrent complaints (Scale: 0 = no, 1 = yes). This index is constructed as follows: I orient all variables so that higher values indicate worse physical health outcomes. Then, I standardize these variables using means and standard deviations from the preperiod. Next, I calculate an equally weighted average of the index components and finally I standardize the final index.
Individual characteristics	
First-generation immigrant	Origin; Scale: 1 = first generation foreign, western background, first generation foreign, non-western background, 0 = dutch, second generation foreign, western background, second generation foreign, non-western background.
Income	Net household income in Euros; Scale: 1 = below the median, 0 = above the median (average pre-period)
Practicing sport activities	Question:"Do you practice sports?"; Scale: 1 = yes, 0 = no (average pre-period)

Variable	Description
Urbanization	Level of urbanization; Scale: 1 = extremely urban, 2 = very urban, 3 = moderately urban, 4 = slightly urban, 5 = not urban (average pre-period).
Index Individual Characteristics	Index sums the variables defined above. As an additional moderator to study heterogeneous treatment effects, I consider whether a the median value of the index of individual characteristics or below the median value.
Family characteristics	
Poor relationship - Mother	Question:"How would you describe your overall relationship with your mother?"; Scale: 1 = not so good, 2 = fairly good, 3 = good, 4 = very good (average pre-period).
Poor relationship - Father	Question:"How would you describe your overall relationship with your father?"; Scale: same as above (average pre-period).
Poor relationship - Family	Question:"How would you describe your overall relationship with your family?"; Scale: same as above (average pre-period).
Parents are separated	Question:"The household head lives together with a partner (wedded or unwedded)"; Scale: 1 = yes, 0 = no (average pre-period).
Index Poor Relationship - Family	Index sums the variables defined above. As an additional moderator to study heterogeneous treatment effects, I consider whether a the median value of the index or below the median value.
Other variabiles	
Pew Research Center	
Worse about your own life	Question:"In general, does social media make you feel? Worse about your own life because of what you see from other friends on social media?"; Scale: 1 = no, 2 = yes, little, 3 = yes, lot.
Pressure to post content that makes you look good	Question:"In general, does social media make you feel? Pressure to only post content that makes you look good to others?"; Scale: same as above.
Pressure to post content that will get likes and comments	Question:"In general, does social media make you feel? Pressure to post content that will be popular and get lots of comments or likes?"; Scale: same as above.
Index Negative Social Media Consequences	The index is constructed aggregating the variables above following the same procedure as the Index Poor Mental Health.
White, non-Hispanic	Race/ethnicity; Scale: 1 = white, non-Hispanic, 0 = otherwise
Black, non-Hispanic	Race/ethnicity; Scale: 1 = black, non-Hispanic, 0 = otherwise
Other, non-Hispanic	Race/ethnicity; Scale: 1 = other, non-Hispanic, 0 = otherwise
Hispanic	Race/ethnicity; Scale: 1 = Hispanic, 0 = otherwise
Income	Coding: 1 = above the median income; 0 = below the median income.

Variable	Description
Region	4-level region; Scale: 1 = Northeast, 2 = Midwest, 3 = South, 4 = West.
Gender	Respondent gender; Scale: 1 = male, 2 = female, 99 = refused.
Statistics Netherlands	
Antidepressants	Percentage of individuals aged 15-25 years to whom in the year concerned antidepressants were dispensed for which the costs are reimbursed under the statutory basic medical insurance.
Antipsychotics	Percentage of individuals aged 15-25 years to whom in the year concerned antipsychotics were dispensed for which the costs are reimbursed under the statutory basic medical insurance.
Anxiolytics	Percentage of individuals aged 15-25 years to whom in the year concerned anxiolytics were dispensed for which the costs are reimbursed under the statutory basic medical insurance.
Mental health (MHI-5)	This is an international standard for a specific measuring of poor mental health, consisting of 5 questions. MHI-5 is actually an extract of "Short Format 36" (SF-36), an elaborate international standard for measuring health. MHI-5 deals with questions related to how one felt during the last 4 weeks. The following questions were asked: (1) Did you feel very nervous? (2) Were you so down in the dump that nothing could cheer you up? (3) Did you feel calm and quiet? (4) Did you feel depressed and down? (5) Were you happy?; Scale: all the time, most of the time, often, sometimes, rarely, and never. The answer categories in positively worded questions of the MHI questionnaire (questions 3 and 5) have been consequently awarded the values 5, 4, 3, 2, 1 and 0. The answer categories in negatively worded questions (questions 1, 2 ad 4) have been awarded the turned-down values. Next, per person the sum scores have been calculated and multiplied by 4, so that the minimum sum score of a person can be 0 (very unhealthy) and the maximum score 100 (perfectly healthy). A score of 60 or more means that a respondent has no poor mental health. A score of less than 60 means that a person does have poor mental health.