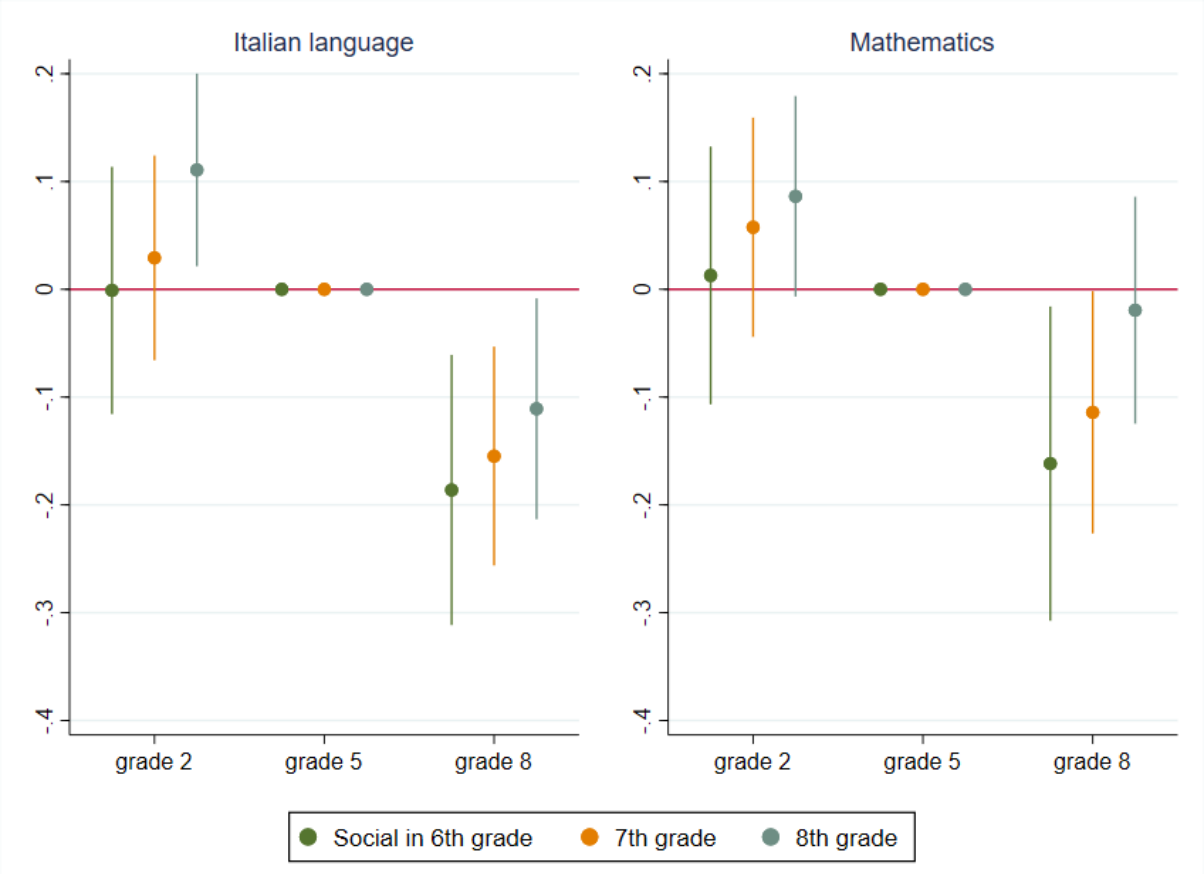
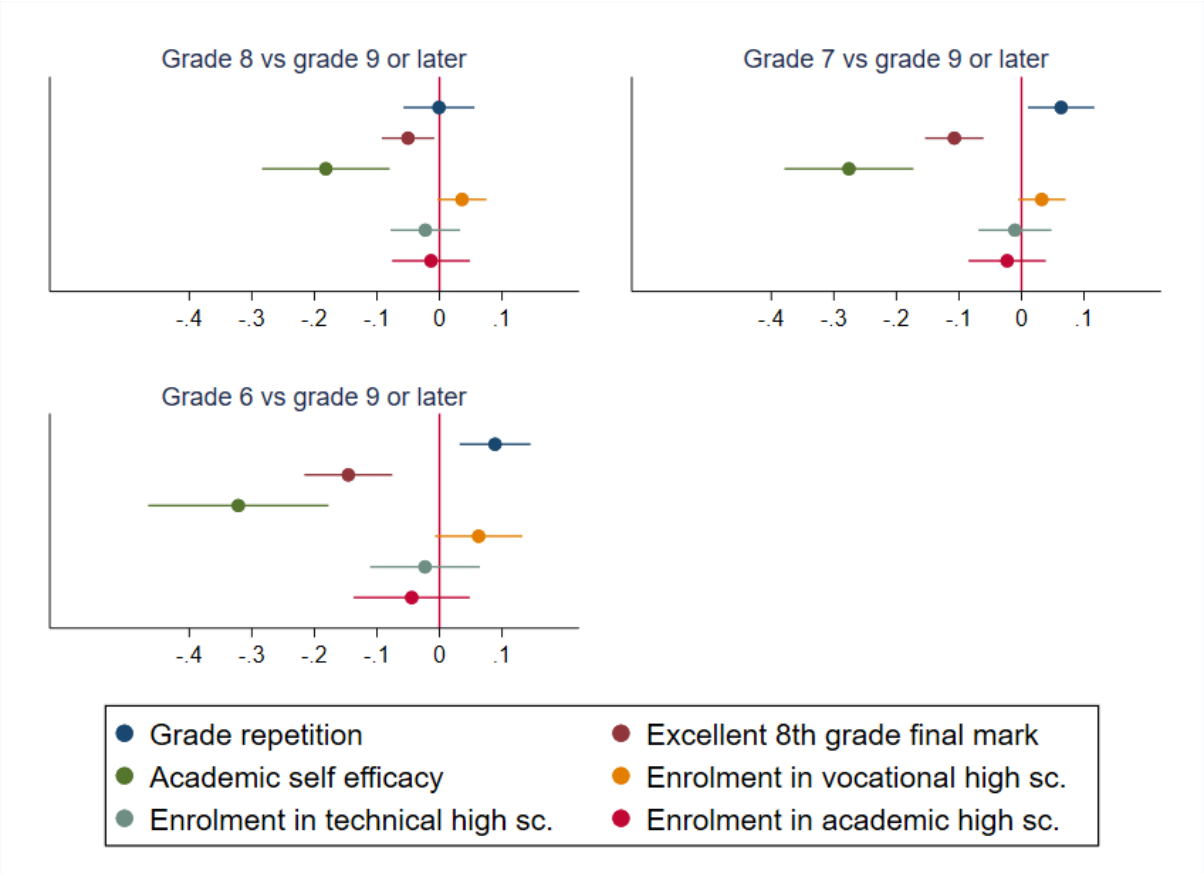
Appendix 1 – Additional tables and figures

Figure A1 – Effect of early use of social networks on school marks in grade 8



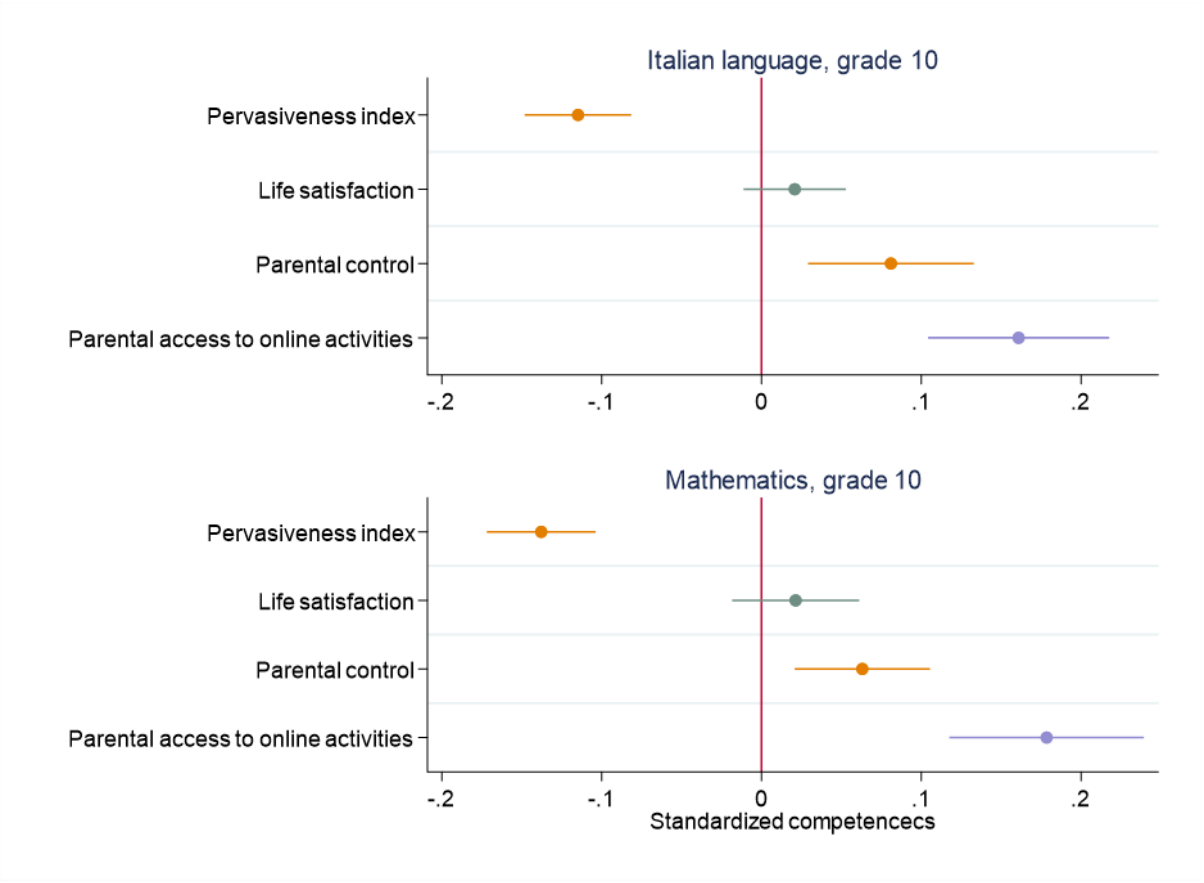
Note: The coefficients display the interaction parameters between treatment status and grade in a student fixed-effects model. Matching method: entropy balancing. Errors clustered at the individual level. N in models comparing the 6th grade early users group to the late users group: 5,458 observations, 1,935 students (mathematics); 5,461 observations, 1,936 students (Italian language). N in models comparing the 7th grade early users group to the late users group: 5,004 observations, 1,771 students (mathematics); 5,008 observations, 1,772 students (Italian language). N in models comparing the 8th grade early users group to the late users group: 4,235 observations, 1,497 students (mathematics); 4,238 observations, 1,498 students (Italian language).

Figure A2 – Effect of early use of social networks on additional academic outcomes



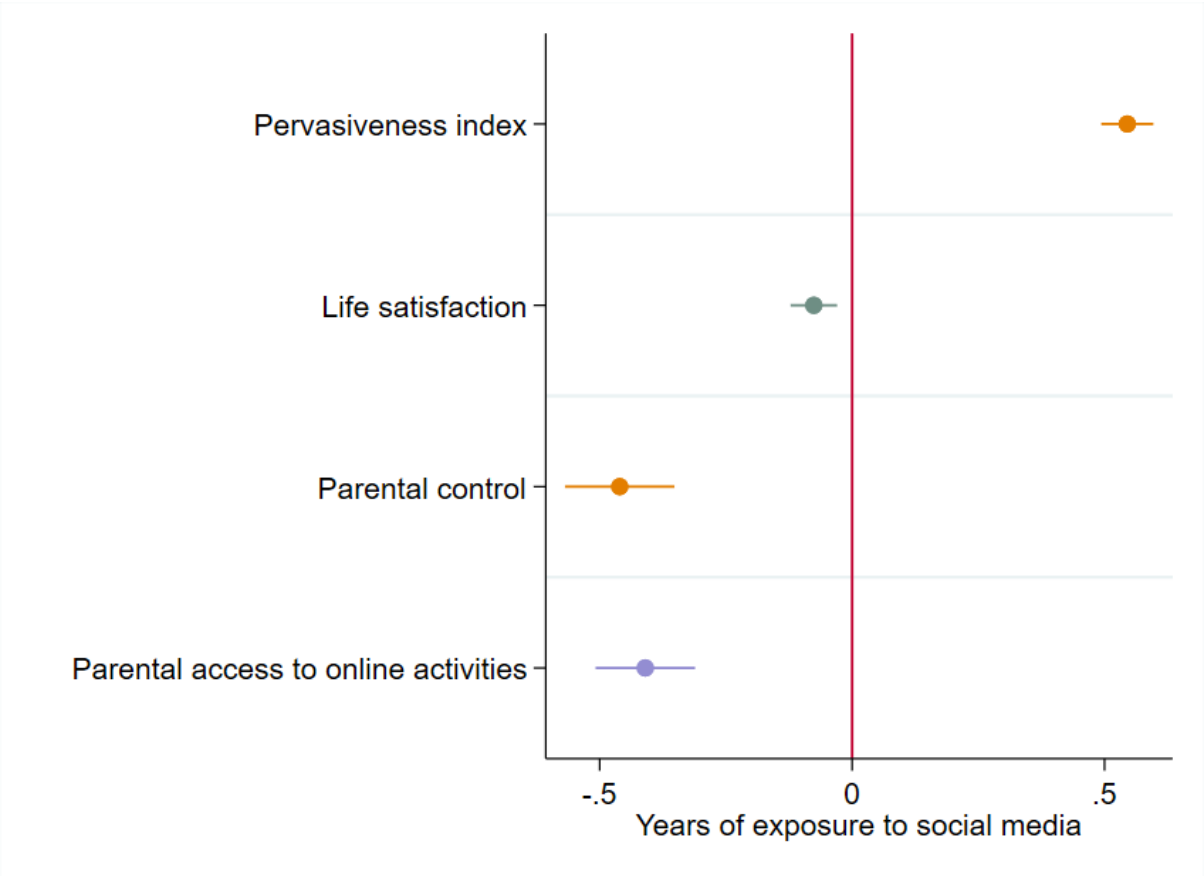
Note: Entropy balancing weights used. Errors clustered at the school level. Outcomes expressed in terms of probability except academic self-efficacy, which is expressed in terms of standard deviations. N for models on marks, self-efficacy and enrolment: 1,956 for models on 6th graders, 1,797 for models on 7th graders, 1,511 for models on 8th graders. N for models on grade repetition (estimated on the 2007 cohort only): 965 for models on 6th graders, 870 for models on 7th graders, 723 for models on 8th graders.

Figure A3 – Association between potential mediators and academic competences in grade 10



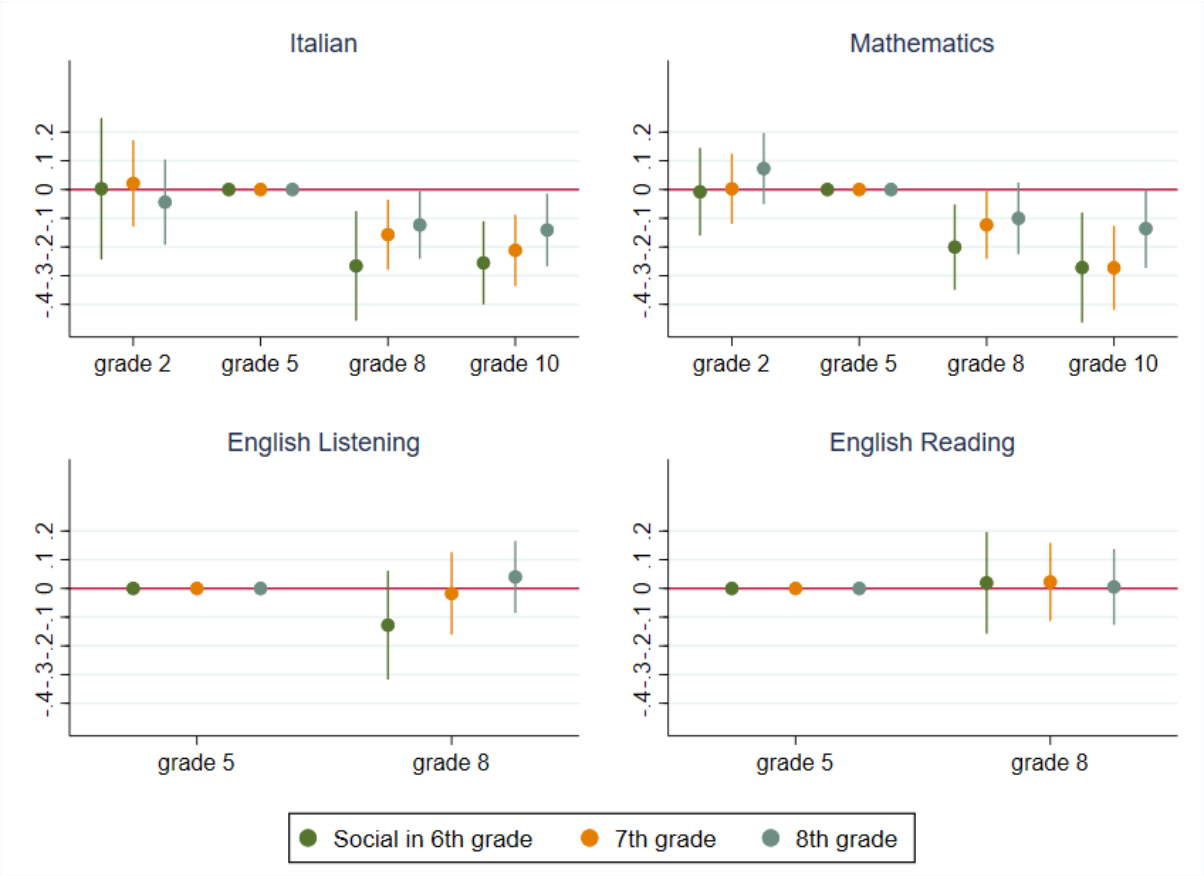
Note: errors clustered at the school level; N: 4,383 for the model on mathematics and 4,414 for the model on Italian. Models estimated separately for each outcome and potential mediator.

Figure A4 – Association between potential mediators and years of exposure to social media by grade 10



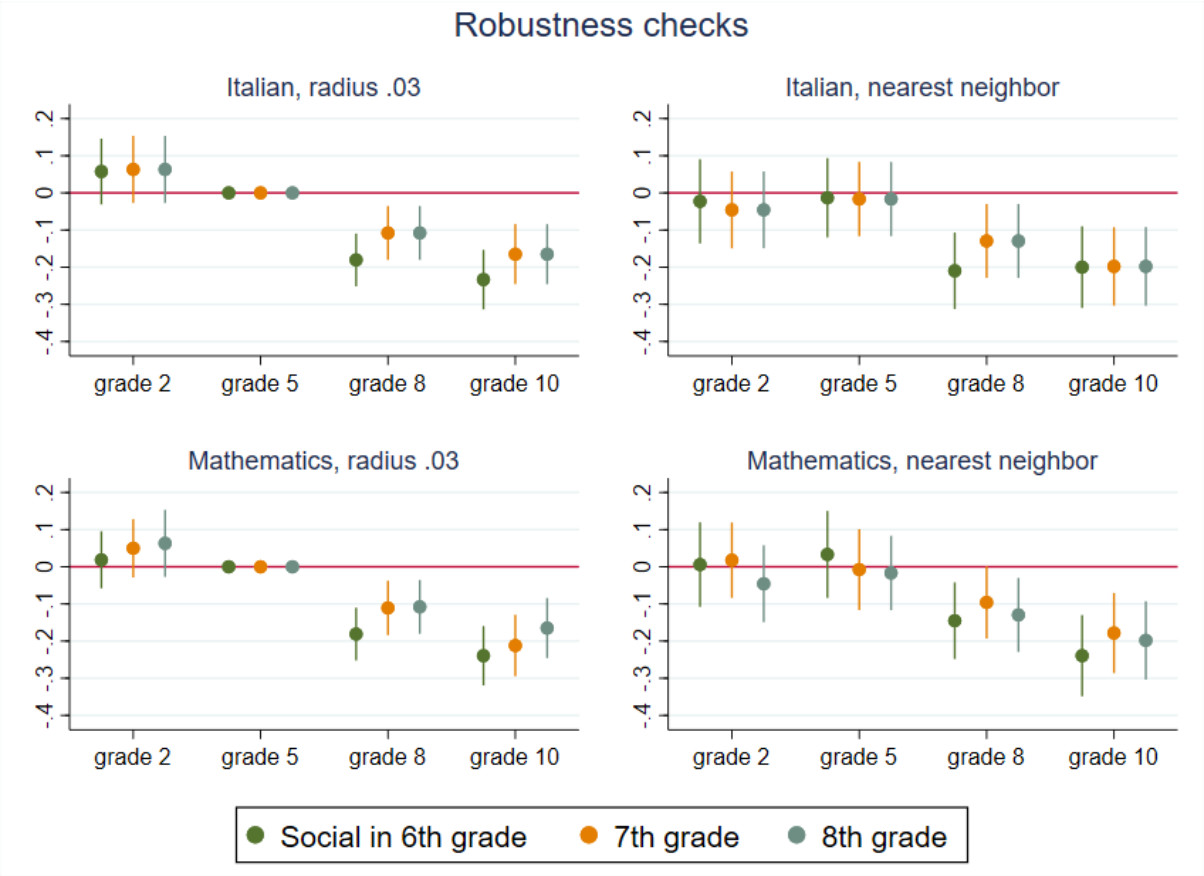
Note: errors clustered at the school level; N: 4,589. Models estimated separately for each outcome and potential mediator.

Figure A5 – Effect of early use of social networks on competences in grades 8 and 10. 2007 cohort only.



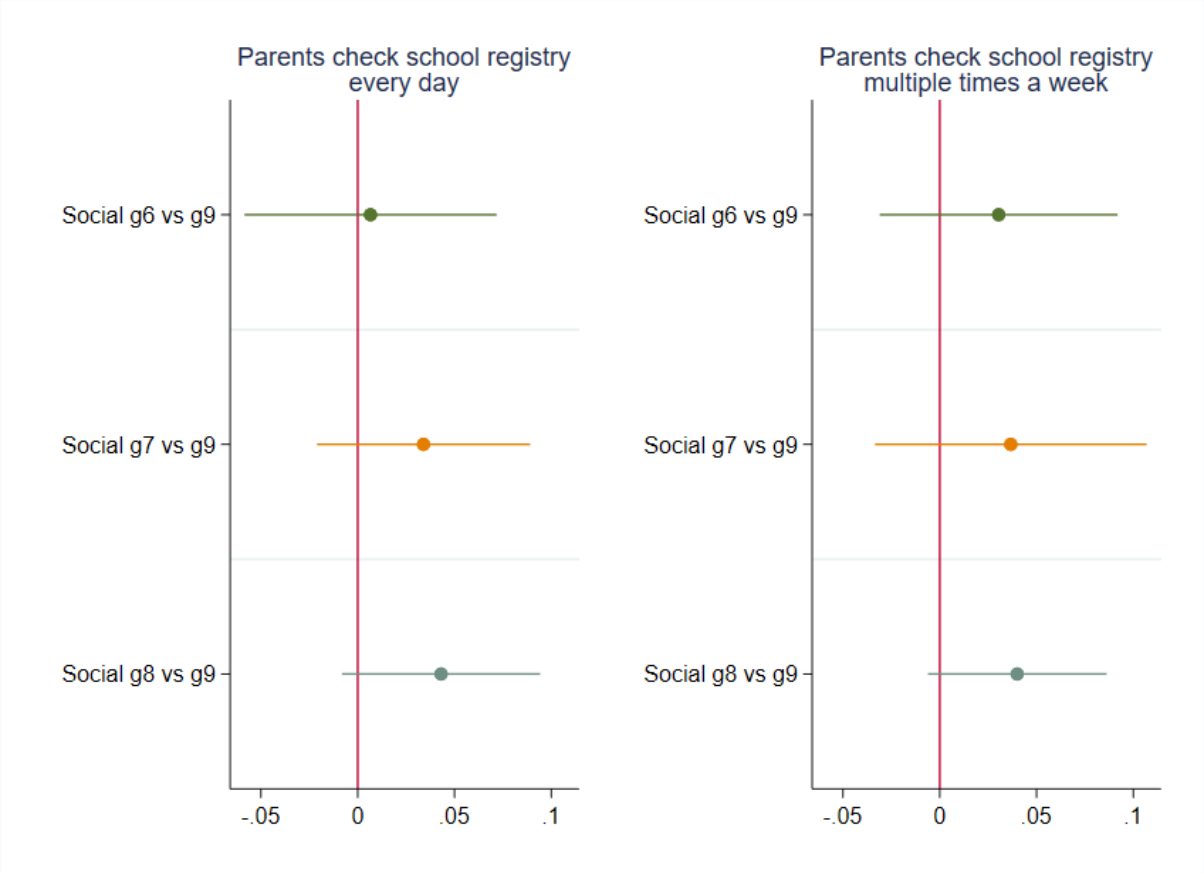
Note: The coefficients display the interaction parameters between treatment status and grade in a student fixed-effects model. Matching method: entropy balancing. Errors clustered at the individual level. N in models comparing the 6th grade early users group with the late users group: 3,846 observations, 990 students (mathematics and Italian); 1,938 observations, 969 students (English reading and listening). N in models comparing the 7th grade early users group with the late users group: 3,408 observations, 880 students (mathematics and Italian); 1,732 observations, 866 students (English reading and listening). N in models comparing the 8th grade early users group with the late users group: 2,828 observations, 728 students (mathematics and Italian); 1,432 observations, 716 students (English reading and listening).

Figure A6 – Effect of early use of social networks on competences in grades 8 and 10, alternative specifications



Note: The coefficients display the interaction parameters between treatment status and grade. Models using radius matching: student fixed-effects models. Models using nearest neighbor matching: matched-pairs fixed-effects models. Errors clustered at the individual level. N in models comparing the 6th grade early users group with the late users group: radius matching 7,570 observations, 1,947 students; nearest neighbor matching 4,157 observations, 1,068 students. N in models comparing the 7th grade early users group with the late users group: radius matching 6,955 observations, 1,794 students; nearest neighbor matching 4,827 observations, 1,240 students. N in models comparing the 8th grade early users group with the late users group: radius matching 5,874 observations, 1,507 students; nearest neighbor matching 4,911 observations, 1,258 students.

Figure A7 – Equivalence between groups in parental frequency of access to online school register; matching technique: entropy balancing.



Note: errors clustered at the school level. N for models on marks, self-efficacy and enrolment: 1,956 for models on 6th graders, 1,797 for models on 7th graders, and 1,511 for models on 8th graders.

Table A1 - Sample representativeness

Track	All selected provinces	EYES UP starting sample	EYES UP final sample
Academic	51	47	49
Technical	33	33	33
Vocational	16	20	18
Total	100	100	100
Total (N students)	99,815	7,083	6,150

Note: we had to exclude students enrolled in three-year vocational programs courses (459 students in our final sample), that are not present in national registers from this table (see footnote 17 for further clarification). Starting sample: students enrolled in the sampled classrooms. Final sample: respondents.

Table A2 - Congruence between effect estimated on different samples

Comparison	Effect estimated at grade	Estimated effect	Difference in the effects against "late users"	Difference in point estimates
(1)	(2)	(3)	(4)	(5)
		Italian		
7th vs 8th grade	8	-0,040	-0,034	-0,006
	10	-0,092	-0,060	-0,032
6th vs 8th grade	8	-0,130	-0,116	-0,014
	10	-0,188	-0,115	-0,073
6th vs 7th grade	8	-0,090	-0,082	-0,008
	10	-0,059	-0,055	-0,004
		Mathematics		
7th vs 8th grade	8	-0,032	-0,026	-0,007
	10	-0,095	-0,104	0,009
6th vs 8th grade	8	-0,109	-0,105	-0,004
	10	-0,125	-0,162	0,037
6th vs 7th grade	8	-0,081	-0,079	-0,001
	10	-0,042	-0,058	0,016

Note: all effects estimated using entropy balancing matching by means of student fixed-effect models. Column 3 reports the effects of the comparison between the groups described in column 1, at the grade specified in column 2. Column 4 reports the difference between the effects estimated against late users of the two groups described in column1 (e.g. the effect of 8th grade against late users minus the effect of 7th grade against late users). Column 5 is calculated by subtracting column 3 from column 4 and highlights the similarity of results by means of different and independent estimation methods.

Appendix 2 – Scales

2.1 Smartphone Pervasiveness Scale for Adolescents (SPS-A)

2.1.1 Literature

Smartphone pervasiveness in students' daily lives was measured through the extended Smartphone Pervasiveness Scale for Adolescence (SPS-A), which was first validated on a sample of 3,289 Italian upper secondary school students (Gerosa et al., 2022). The scale assesses the subjective frequency of smartphone use in key daily-life moments that could affect adolescents' social and physiological well-functioning (e.g. during mealtimes with family, while spending time with friends and on waking up). When assessing non-pathological problematic smartphone use, the SPS-A has been proved to be both a valuable alternative to smartphone addiction scales and a good predictor of smartphone duration of use with respect to self-reported measures (Chakraborty et al., 2024). Students were asked to report on a five-point Likert scale ranging from "Never" to "Always" (*Never = 1, Rarely = 2, Sometimes = 3, Often = 4, Always = 5*) how frequently they usually make use of their smartphones during seven key moments of the day (see *Table A3*).

2.1.2 Reliability and construct validity

In terms of inter-item reliability, the items included in the analysis displayed an acceptable degree of internal consistency (Cronbach's alpha = 0.74). Corrected item-total correlations, i.e. the relationship between single items and the overall scale when the concerned item is removed, were investigated: the values were all positive and above the 0.3 threshold.
Construct validity was assessed by means of a Confirmatory Factor Analysis (CFA) on the entire sample, verifying the unidimensionality of the latent construct, i.e., Smartphone Pervasiveness. The Weighted Least Square Mean and Variance estimation method (WLSMV) was adopted, being one of the best performing estimators when dealing with ordinal data (Brown, 2006). In accordance with the rules of thumb provided in the literature (Hu & Bentler, 1999, Browne and Cudek, 1993, MacCallum et al., 1996) the model specification results in an acceptable fit (RMSEA = 0.072 [0.068-0.076], SRMR = 0.047, CFI = 0.936 , TLI = 0.915). The standardized factor loadings were all positive, significant and above the 0.4 threshold (Brown, 2006) (see Table A4 for further details).

2.1.3 Items and computation

The pervasiveness index is computed by taking the mean of the nine items composing the SPS-A scale.

Table A3 - Smartphone Pervasiveness Scale for adolescents (SPS-A); *Italian version in brackets*

How often do you find yourself using your smartphone during the following everyday situations? (*Quanto spesso ti capita di usare lo smartphone durante le seguenti situazioni quotidiane?*)
1 - During meals with my family (*Durante i pasti con i miei familiari*)
2 - While I'm with friends (*Mentre sono con gli amici*)
3 - While doing homework (*Durante lo svolgimento dei compiti a casa*)
4* - The evening before falling asleep (*La sera prima di prendere sonno*)
5 - At night if I wake up (*Di notte se mi sveglio*)
6 - In the morning as soon as I wake up (*La mattina appena mi sveglio*)
7 - During lessons at school, including when we are not supposed to be using devices (*Durante le lezioni a scuola, anche quando non è previsto l'utilizzo di dispositivi*)
8 - While watching a series, a film or TV (*Mentre guardo una serie, un film o la TV*)
9* - While doing physical activities (e.g. sports, trips, housework) (*Mentre svolgo attività di movimento (es. sport, gite, lavori in casa)*)

*Items added to the SPS-A scale.

Table A4 - Descriptive Statistics, factor loadings of the one-factor CFA model and factorial validity of the SPS-A latent construct.

Item	**n**	**Mean**	**Std dev**	**Skewness**	**Excess kurtosis**	**Factor Loadings**
1	6609	1.877	1.074	1.148	0.527	0.475
2	6609	3.03	0.912	0.047	-0.338	0.470
3	6609	3.299	1.031	-0.284	-0.352	0.531
4	6609	4.271	1.072	-1.571	1.744	0.676
5	6609	2.234	1.427	0.830	-0.713	0.664
6	6609	3.462	1.416	-0.381	-1.214	0.599
7	6609	2.664	1.174	0.200	-0.811	0.489
8	6609	2.921	1.179	0.007	-0.875	0.530
9	6609	2.037	1.164	0.922	-0.136	0.444
Fit Indices (WLSMV estimation method) – one factor CFA						
2	df	p-value	RMSEA	CFI	TLI	SRMR
958.673	27	0.000	0.072 [0.068-0.076]	0.936	0.915	0.047

2.2 Subjective well-being

2.2.1 Literature

The Subjective well-being scale was first developed to assess one of the seven core theoretical dimensions of psychological well-being, measured by means of the Comprehensive Inventory of Thriving for Children (CTI) (Su et al., 2014). Subjective well-being has been defined as "an internal barometer of how life is going" and it is articulated in 3 facets, i.e. life satisfaction, positive feelings and negative feelings. Each subscale is defined by 3 indicators answered on a 5 point-Likert scale ranging from "Strongly disagree" to "Strongly agree" (*Strongly disagree = 1, Disagree = 2, Neither agree nor disagree = 3, Agree = 4, Strongly agree = 5*) (see Table A5). The psychometric features of the Italian adaptation of the CIT have been evaluated in previous research (Andolfi et al., 2017), providing evidence of the validity and reliability of the instrument.

2.2.2 Reliability and construct validity

Concerning inter-item reliability, the items included in the analysis displayed a good/excellent degree of internal consistency with respect to both the total SWB scale (Cronbach's alpha = 0.91) and the single subscales (Life satisfaction Cronbach's alpha = 0.85, Positive feelings Cronbach's alpha = 0.90, Negative feelings Cronbach's alpha = 0.83). Corrected item-total correlations values with respect to both the total scale and the subscales were all positive and above the 0.3 threshold.

The three facets of psychological well-being are assumed to be unidimensional and distinguishable from each other: a CFA was conducted on the whole sample to test the posited factor structure. A 3-factor model was specified: single items loaded on the hypothesized latent variable and factors were allowed to correlate with each other. The Weighted Least Square Mean and Variance estimation method (WLSMV) was adopted, being one of the best performing estimators when dealing with ordinal data (Brown, 2006). Missing values were treated through listwise deletion, i.e., removing individual records with missing values from the analysis. In accordance with the rules of thumb provided in the literature (Hu & Bentler, 1999, Browne and Cudek, 1993, MacCallum et al., 1996) the model specification results in an acceptable fit (RMSEA = 0.065 [0.061-0.069], SRMR = 0.019, CFI = 0.995 , TLI = 0.993). The standardized factor loadings were all positive, significant and above the 0.4 threshold (Brown, 2006) (see Table A6). Factor correlations were all below the 0.8 threshold ([0.62 – 0.79] range), supporting the notion that the three subscales represent related yet distinguishable latent constructs (Brown, 2006).

2.2.3 Scale computation

As reported in Appendix A on the website of the first author of the CIT development and validation paper, the score on each subscale is computed by averaging the responses across three items on the scale.

Table A5 - Subjective well-being; *Italian version in brackets*

Please indicate your agreement or disagreement with each of the following statements using the scale below (1 Strongly Disagree, 2 Disagree, 3 Neither Agree nor Disagree, 4 Agree, 5 Strongly Agree). (*Indica quanto sei d'accordo con ciascuna delle seguenti frasi, selezionando la tua scelta sulla seguente scala che va da 1 "Per niente d'accordo" a "Molto d'accordo".*)
1 - In most ways my life is close to my ideal (*In molte cose la mia vita è come la vorrei*)
2 - I am satisfied with my life (*Sono contento/a della mia vita*)
3 - My life is going well (*La mia vita va bene*)
4 - I feel positive most of the time (*La maggior parte delle volte mi sento contento/a*)
5 - I feel happy most of the time (*La maggior parte delle volte mi sento felice*)
6 - I feel good most of the time (*La maggior parte delle volte mi sento di buon umore*)
7 - I feel negative most of the time (*La maggior parte delle volte mi sento triste*)
8 - I experience unhappy feelings most of the time (*La maggior parte delle volte mi sento infelice*)
9 - I feel bad most of the time (*La maggior parte delle volte mi sento di cattivo umore*)

Table A6. Descriptive Statistics, factor loadings of the 3-factor CFA model.

Item	**n**	**Mean**	**Std dev**	**Skewness**	**Excess kurtosis**	**Subscale**	**Factor Loadings**
1	6599	3.334	1.050	-0.396	-0.516	Life satisfaction	0.750
2	6587	3.653	1.057	-0.703	-0.064		0.922
3	6551	3.664	1.051	-0.697	-0.031		0.887
4	6585	3.473	1.037	-0.429	-0.375	Positive feelings	0.928
5	6584	3.422	1.046	-0.373	-0.44		0.939
6	6588	3.355	1.035	-0.324	-0.455		0.841
7	6581	3.322	1.037	-0.344	-0.475	Negative feelings	0.867
8	6577	3.488	1.098	-0.424	-0.549		0.889
9	6601	3.195	1.144	-0.234	-0.749		0.748
Fit Indices (WLSMV estimation method) – three factors CFA							
2	df	p-value	RMSEA	CFI	TLI		SRMR
674.946	24	0.000	0.065 [0.061-0.069]	0.995	0.993		0.019

2.3.1 Literature

Perceived self-efficacy is defined as the "belief in one's capabilities to organize and execute the courses of action required to produce given attainments (Bandura, 1997). Self-efficacy is supposed to play a general role on mental health, determining how people feel, think, motivate themselves and behave and therefore sustaining personal accomplishments, reducing stress and lowering vulnerability to depression and anxiety (Tahmassian & Jalali Moghadam, 2011, Muris, 2002). It has been posited that three different domains of self-efficacy are involved in the regulation of negative affect: academic, social and emotional self-efficacy (Muris, 2001). Academic self-efficacy refers to the individual's perceived ability to control their learning behaviours, master subjects and meet educational expectations. Social self-efficacy refers to the individual's perceived ability to be authentic and assertive in peer relationships. Emotional self-efficacy refers to the individual's perceived ability to cope with negative emotions (Muris, 2001). The self-efficacy questionnaire for children was developed to account for the above-mentioned domains, each of which was originally defined by eight items. Due to optimization reasons related to the length of the questionnaire, not all the original items were adopted in the questionnaire. First, one item per subscale was removed considering Muris's (2001) EFA, which provided evidence that some of the items considered did not indeed load convincingly on their intended factor. Second, of the seven remaining items per domain, four were selected and implemented in the questionnaire. The choice relied on the combination of three main criteria: i) research objectives, ii) avoidance of overlapping items, i.e., items similar in meaning items and iii) the higher factor loadings value according to Muris (2001).
The final 12-item scale was measured on a 5-point Likert scale, with values ranging from "Not at all" to "Very well" (*Not at all = 1, Very well = 5*).

2.3.2 Reliability and construct validity

Analysis of the inter-item reliability of subscales suggested removing one item from the social self-efficacy subscale and one item from the emotional self-efficacy subscale, as the two items were both associated with an increase in reliability following their respective drop. To produce further evidence supporting the removal of the two items, an exploratory analysis was conducted to investigate item clustering in greater detail. The sample was randomly divided into two halves: EFA was conducted on the first half, while CFA was then implemented on the second half. According to the literature, parallel analysis suggested the retention of three factors, which were extracted by means of Principal Axis Factoring and obliquely rotated (oblimin rotation). As both the items concerned did not load convincingly on the intended factor, the research team agreed on their removal. Following the above-mentioned adjustment, the indicators displayed acceptable/good degree of internal consistency with respect to both the total self-efficacy scale (Cronbach's alpha = 0.71) and the single subscales (academic self-efficacy Cronbach's alpha = 0.74, social self-efficacy Cronbach's alpha = 0.63, emotional self-efficacy Cronbach's alpha = 0.78). Corrected item-total correlations values with respect to both the total scale and the subscales were all positive and above the 0.3 threshold.
The three self-efficacy domains are assumed to be unidimensional and distinguishable from each other: a CFA was conducted on the whole sample to test the posited factor structure. A 3-factor model was specified: single items loaded on the hypothesized latent variable and

factors were allowed to correlate with each other. The Weighted Least Square Mean and Variance estimation method (WLSMV) was adopted, being one of the best performing estimators when dealing with ordinal data (Brown, 2006). In accordance with the rules of thumb provided in the literature (Hu & Bentler, 1999, Browne and Cudek, 1993, MacCallum et al., 1996) the model specification results in an acceptable fit (RMSEA = 0.056 [0.052-0.060], SRMR = 0.035, CFI = 0.977 , TLI = 0.968). The standardized factor loadings were all positive, significant and above the 0.4 threshold (Brown, 2006) (see *Table.A8* for further details). Factor correlations were all far below the 0.8 threshold ([0.12 – 0.51] range), supporting the notion that the three subscales represent distinguishable latent constructs (Brown, 2006).

2.3.3 Scale computation

The score on each subscale is computed by averaging the responses across the items on the scale.

Table A7 - Self-efficacy; *Italian version in brackets*

How well can you… (1 - "Not at all", 5 - "Very well") ***(Indica con un valore compreso tra 1 "Per niente" e 5 "Molto", scegliendo la risposta appropriata per ciascun elemento. Quanto riesci a…)***
1 – Can you study when there are other interesting things to do? (*Studiare quanto ci sono altre cose interessanti da fare?*)
2 – Can you express your opinions when other classmates disagree with you? (*Esprimere le tue opinioni quando altri compagni di classe non sono d'accordo con te?*)
3 – Do you succeed in cheering yourself up when an unpleasant event has happened? (*Tirarti su di morale quando è accaduto un evento spiacevole?*)
4 – Can you study a chapter for a test? (*Studiare un capitolo per una verifica-interrogazione?*)
5 – Can you become friends with other children? (*Fare amicizia con gli altri coetanei?*)
6 – Do you succeed in suppressing unpleasant thoughts? (*Scacciare pensieri spiacevoli?*)
7 – Do you succeed in finishing all your homework every day? (*Finire tutti i tuoi compiti ogni giorno?*)
9 – Can you give yourself a pep talk when you feel low (*Darti una carica positiva quando ti senti giù?*)
10 – Can you pay attention during every class? (*Prestare attenzione durante ogni lezione?*)
11 – Can you tell other children that they are doing something that you don't like? (*Dire ai tuoi coetanei che stanno facendo qualcosa che non ti piace?*)

Removed items:
Social self-efficacy subscale: 8 – Can you work in harmony with your classmates? (*Lavorare in armonia con i tuoi compagni di classe?*)
Emotional self-efficacy subscale: 12 – Do you succeed in not worrying about things that might happen? (*Vivere le cose che potrebbero accadere senza preoccupazione?*)

Table A8 - Descriptive Statistics, factor loadings of the 3-factor CFA model.

Item	n	Mean	Std dev	Skewness	Excess kurtosis	Subscale	Factor Loadings
1	6609	2.47	1.06	0.37	-0.43		0.643
4	6609	3.47	1.16	-0.44	-0.62	Academic	0.717
7	6609	3.00	1.34	0.00	-1.19	self-efficacy	0.705
10	6609	3.03	1.06	-0.06	-0.60		0.655
2	6609	3.30	1.27	-0.22	-1.00		0.666
5	6609	3.41	1.15	-0.32	-0.71	Social self-efficacy	0.605
11	6609	3.25	1.21	-0.20	-0.88		0.659
3	6609	2.98	1.21	0.04	-0.91		0.759
6	6609	2.71	1.18	0.27	-0.78	Emotional self-efficacy	0.684
9	6609	2.95	1.20	0.06	-0.86		0.856
Fit Indices (WLSMV estimation method) – three factors CFA							
2	df	p-value	RMSEA	CFI	TLI	SRMR	
695.467	32	0.000	0.056 [0.052-0.060]	0.977	0.968	0.035	