RESEARCH

Supplemental Information – Mapping Language Literacy At Scale: A Case Study on Facebook

Yu-Ru Lin[1*], Shaomei Wu[2] and Winter Mason[3]

[*]Correspondence: yurulin@pitt.edu
[1]School of Computing and Information, University of Pittsburgh, Pittsburgh, PA, USA
Full list of author information is available at the end of the article

S1 Supplementary Text

S1.1 Language literacy and visual information consumption

We examine the relationship between populations' language literacy levels and their interest in different types of content. Because our assessment concerns the ability to process textual content, we assume there exists a negative relationship between a population's literacy estimate and their attention toward non-textual (e.g., visual) content. Fig. S4 shows the correlations between countries' `OLLE`'s (x-axes) and the relative time spent on visual content by the countries' Facebook population (y-axes). As expected, populations with a lower level of language literacy tend to spend relatively more time on visual content. The global correlation is -0.38 (Spearman's rank correlation; $p < 0.001$), with correlations over different areas ranging from -0.29 (Latin America & the Caribbean; $p < 0.062$) to -0.62 (Europe/Oceania/Northern America; $p < 0.001$)[1].

S1.2 Elbow range detection in popularity curves

LoFF words are determined based on "elbow" range on the word popularity curve for each language, where the relative word frequencies begin to have a systematic decline. Fig. S1 shows the popularity of words in decreasing order, i.e., from the most to the least popular word, as popularity curves. It can be seen that a systematic decline in the word popularity appears at the point of maximum curvature in a popularity curve. In other words, the interest region associated with LoFF words corresponds to the "elbow" (or "knee") point on the smoothed word popularity curve. Mathematically, the curvature is a mathematical measure of how much a function differs from a straight line [49, 50]. Estimating the knee/elbow point for a continuous function is straightforward since the curvature is well-defined for continuous functions; however, it is a challenging task for discrete data. It is also an inherently heuristic process [50]. To reliably detect the elbow range, we leverage the "Kneedle" detection [49], an efficient algorithm that can efficiently detect knee points in discrete data, and the standard maximum curvature approach on a smoothed function learned from the discrete points. First, we employ generalized additive models with cross-validation to learn a smooth function for each of the popularity curves [**?**]. As shown in Fig. 1 E-G, we define an elbow range as an area between two points k_0 and k_1 (highlighted in red) that best describe the systematic

[1] As two of the seven geographical groups only have few countries, we merge the seven groups into five (based on proximity) to provide an adequate statistical description.

decline in the curve. The two points were determined by combining two heuristic methods: (i) the standard maximum curvature points that can be calculated from any continuous function, and (ii) the approximate knee points (Kneedle detection method) based on the notion that knee points differ most from the straight line connecting the curve's two endpoints [49]. Note that while the two notions may be considered to be conceptually similar, the approximate knee points (the second notion) are not necessarily the maximum curvature points especially when the curves are skewed. In a right-skewed curve (as in the case of a word popularity curve), the approximate knee points tend to fall into the right of the maximum curvature points. Thus, we detect an elbow by two points k_0 and k_1 through maximum curvature measurement and approximate knee point detection method respectively. Unlike other knee/elbow detention methods that are sensitive to noises and rescaling, we found this hybrid approach is more robust to rescaling and small fluctuations in our data. Words with ranks falling into the elbow range are considered to be the "LoFF words" in the language. Fig. 1 A-C highlights the elbow range detected from the word popularity curve for each of the three most used languages, and Fig. S1 shows the detected elbow ranges for all 12 languages.

S1.3 Procedure for estimating online language literacy

For a given population with a given language, the procedure to measure the collective language literacy involves the following steps:

(i) Processing of user-generated texts: We use public posts written in any of the chosen languages created by Facebook users who are at least 18 years old and active during a 30-day period between April 20 and May 20, 2020. We exclude posts that did not contain any text or text that was shorter than 2 characters or longer than 1000 characters, as well as posts that contained URLs as these are more likely to be copied and pasted from other sources rather then composed by users.

(ii) Aggregate statistics per user: After tokenizing the public posts, for each user, we quantify the number of unique words (unigrams) that falls in the range of LoFF words. We then obtain a relative LoFF word count w_u that is normalized by the active level of post creation per user, i.e., w_u is given by (the total number of LoFF words observed from u's public posts) / (the total number of u's public posts). We count each unigram once per user, regardless of the frequency used, to avoid overestimating the use of particular words or the inflation from copy-pasted content.

(iii) Aggregate statistics per population: For each geographically bounded community (e.g. a county or a region) with at least 1000 active users observed in the study period, the population-level estimate is calculated as $\bar{w}$, the average of w_u's over all active users u's in the geographically bounded community. The gender- or region-disaggregate population-level estimates also require a minimum of 1000 active users observed in the study period in any of the disaggregate groups. The threshold of 1000 unique users from any group was chosen to ensure user privacy and the statistical power of our method. We also exclude users who produced a high volume of posts (above 75 percentiles) to avoid a small number of highly productive users dominating the measurement.

Throughout the procedure, none of the personal identifiable information or any personal or private content was used. Only the aggregate statistics w_u and $\bar{w}$ were generated from the process.

How to retrieve pre-computed LoFF words LoFF words are determined based on the word popularity curves derived from the Facebook users' use of the up 200,000 most frequent words in each language. These pre-computed LoFF words can be retrieved through the following steps:

(i) Install the fastText[2]

(ii) Run `download_model.py $lang` to get the dictionary of a specific language, where `$lang` is the language indicator (e.g., `en` for English, `es` for Spanish). This script will download the dictionary in a binary file (let `$filename` be the filename of the downloaded file).

(ii) Run `fasttext dump $filename dict > $ofilename` to convert the binary dictionary file to a text file (let `$ofilename` be the filename of the output text file). This file contains up to 200,000 lines where each line is a word and its frequency. The frequencies can be used to rank the words from the most to the least frequent.

(iv) Extract the LoFF words based on the knee points listed in Table S1. For example, the LoFF words in English correspond to the words in the fastText dictionary that are ranked between 5,000 to 9,000 in the decreasing order of word frequency.

S1.4 Case study: India as a multilingual country

We choose India as a case study for countries using multiple languages to study the effect of choosing the most used language as a single representative language for literacy estimate. While India uses Hindi and English as official languages nationwide, it has no single national language. It has over 30 states/union territories, each of which has its own official language(s). There are 22 official languages recognized by country officials, in addition to some other languages recognized as additional official languages at the regional level. In this analysis, we include the additional five most used languages in India according to the India census reported in 2011 [**?**]: Hindi (43.6%), Bengali (8.3%), Marathi (7.1%), Telugu (6.9%), and Tamil (5.9%). Languages with less than 5% speakers among the Indian national population are not considered. On Facebook, the most used language in Indian users' public posts is English (en), which has about three times the users posting in Hindi (hi), and about 20, 44, 90, and 181 times those posting in Bengali (bn), Marathi (mr), Telugu (te), and Tamil (ta), respectively. Across regions, the non-English language using populations on Facebook are sparse. Only 14 (48.2%) regions have more than 10% of the number of English posters posting in Hindi, and only 4 (13.8%) and 1 (3.4%) regions have more than 1% of the number of English posters posting in Bengali and Marathi.

We then create a language literacy estimation for each of the six languages, using the same approach but include the additional languages (Hindi, Bengali, Marathi, Telugu, and Tamil) from fastText unigram data [47]. For validation, we gather the

[2] `https://fasttext.cc/docs/en/python-module.html#installation`

regional literacy survey reported in the India census 2011 [?], which is the most recent data available. Fig. S7 shows the comparison of our language estimation with the officially reported literacy data. We first estimate the language literacy for every language. Fig. S7 A and C-E show the estimation based on posts in a single language. Note that, while there are regional differences, the use of Hindi, Bengali, and Marathi is extremely sparse in most regions. Therefore, the non-English language estimates alone cannot be directly used to create a regional measurement. Due to the sparse use of non-English language on the platform, the correlation between the non-English language estimates and the reported literacy is insignificant. We additionally create a multi-language estimation weighted by the popularity of each language within a region, as shown in Fig S7 B. We observe that both English-only and multi-language estimates (without any additional calibration) have significant correlations with the reported literacy data (Spearman's rank correlations with positive 95% CIs and $p < 0.005$). However, literacy estimation based on multiple languages has neglectable improvement over English-based estimation in terms of the correlations – from 0.51 to 0.52. This is likely due to the low rate of users posting in non-English languages in many regions. Here, the comparison relies on the officially reported literacy data, which has a limitation: they do not capture the change in regional literacy levels since 2011, and likely do not properly reflect the diverse language skills used by minority populations. This case study illustrates the challenge of obtaining gold-standard literacy measures for multilingual countries. While this does not prevent us to create a multi-language literacy measurement per country, for validation purposes, we simply choose a single representative language for multilingual countries. Thus the correlation should be interpreted with caution – the officially reported data that guide this choice often has a bias against language minorities.

S1.5 Robust check: regional disparity and language dominance

Our `OLLE` is created based on a country's representative language, i.e., the language used by the most Facebook users in the country. One potential risk of relying on a single representative language is that the regional disparity measure in a multilingual country may simply capture the distribution of languages, rather than the diversity of language skills. To test this, we examine the relationship between the user percentage of the representative language in a multilingual country and the country's regional disparity measure. Among the 167 countries studied, there are 20 multilingual countries, but only 13 meet the criteria to have a regional disparity measure. Recall in Section S1.3 that each geographically bounded community (i.e., in this case, a region within a country) with at least 1000 active users observed in the study period. For these 13 multilingual countries, we plot the countries' percentage of Facebook users using the representative language on the y-axis and on the x-axis, either (A) `OLLE`, or (B) regional disparity as shown in Fig. S8. If there is a systematic bias, e.g., countries with low representative language user percentages tend to have high a regional disparity measure, we would see a trend in such a plot. However, we do not observe a clear systematic bias. While our sample size is limited, this analysis is helpful for checking whether there is a potential bias in the small sample of multilingual countries.

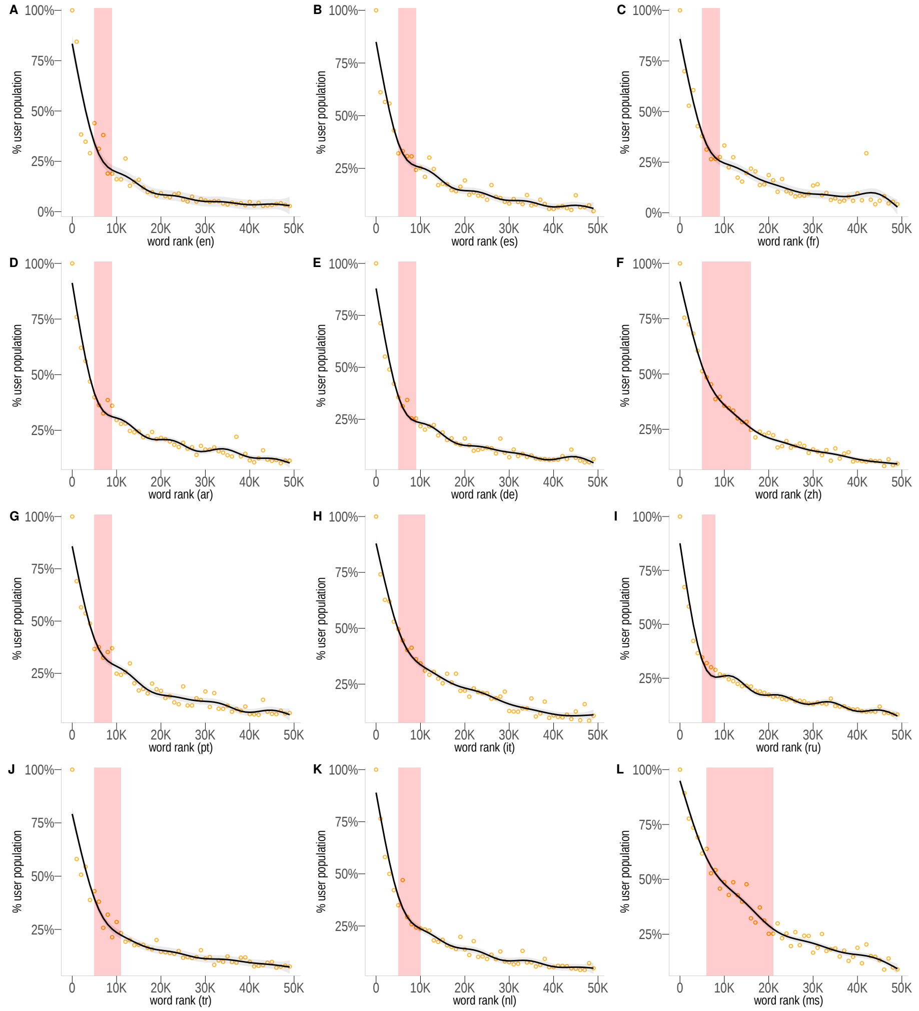

Figure S1: Determining the “LoFF word” range using curvature and knee points detection. In half of the studied languages, the ranks of the big words range between 5000 and 9000. Others (zh, it, ru, tr, nl, and ms) have wider or narrower ranges.

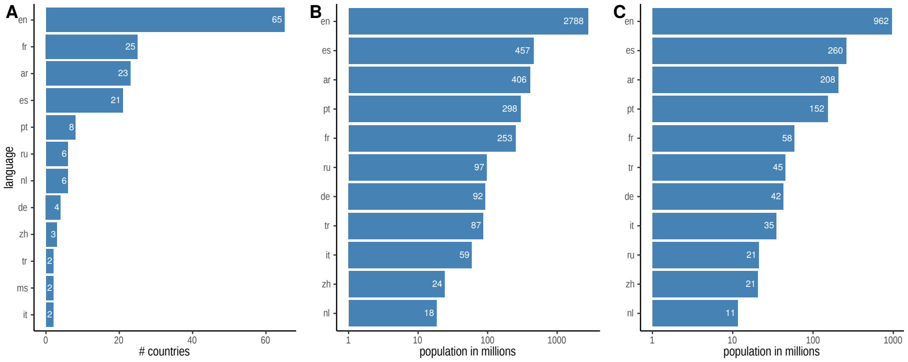

Figure S2: Countries covered in our estimation. There are 167 countries in 12 different languages, including 147 countries with a single or dominant language and 20 multilingual countries. For a multilingual country (having multiple official languages), we use the most used language of the country to estimate its language literacy. (A) The number of countries in each language. (B) The total population in each language. (C) The Facebook user count in each language, according to publicly available information about Facebook penetration statistics in 2022 [**?**]. The populations from the 20 multilingual countries are excluded in (B) and (C) because we do not have the sub-population estimates of different languages within these multilingual countries.

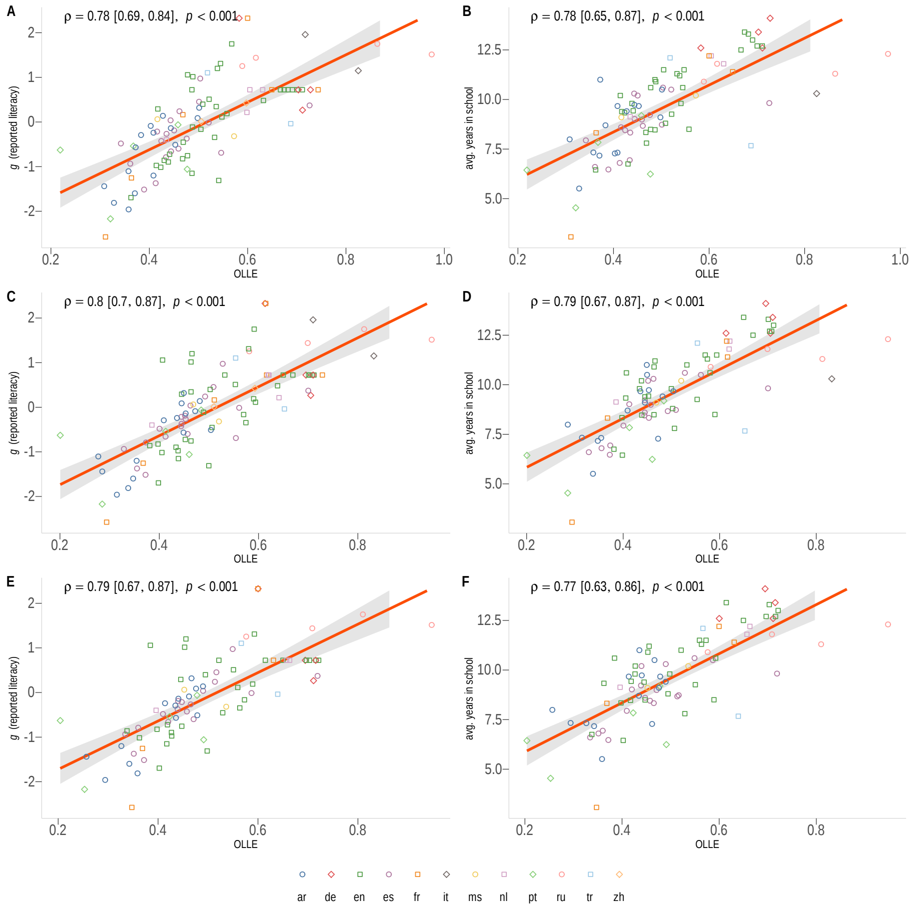

Figure S3: The literacy estimate (x-axis) obtained from Models (a,b,c) listed in Table S2, compared with the reported literacy rate and the education in terms of schooling years (y-axis). The reported literacy rate was transformed for normality. (A,B) Literacy estimate adjusted by Model (a). (C,D) Literacy estimate adjusted by Model (b). (E,F) Literacy estimate adjusted by Model (c).

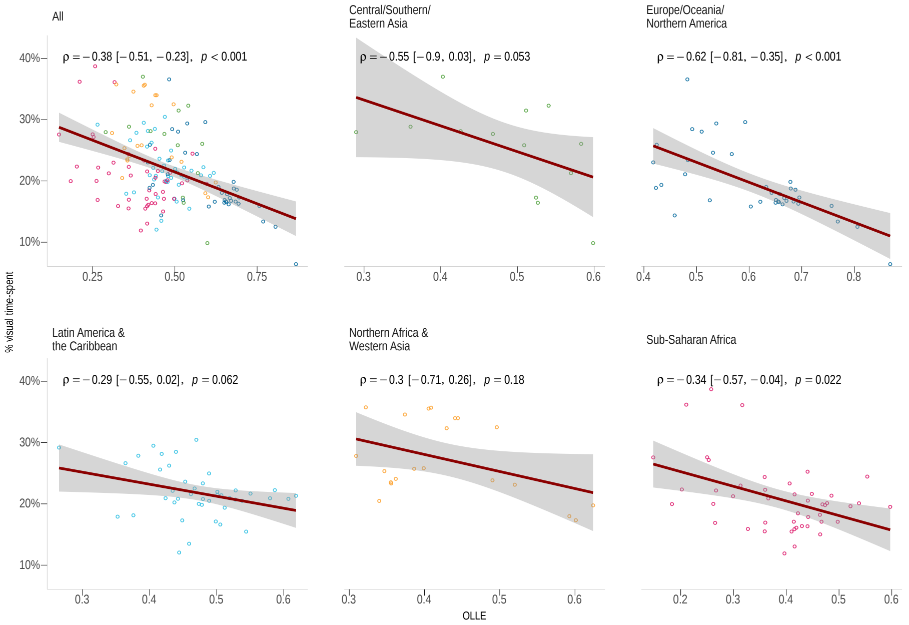

Figure S4: Relationship between countries' OLLE's (x-axes) and the relative visual time-spent (y-axes), where the relative visual time-spent is given as the proportion of time spent on photos and videos relative to the time on news feeds, photos and videos combined. Correlations are reported based on Spearman's rank correlation.

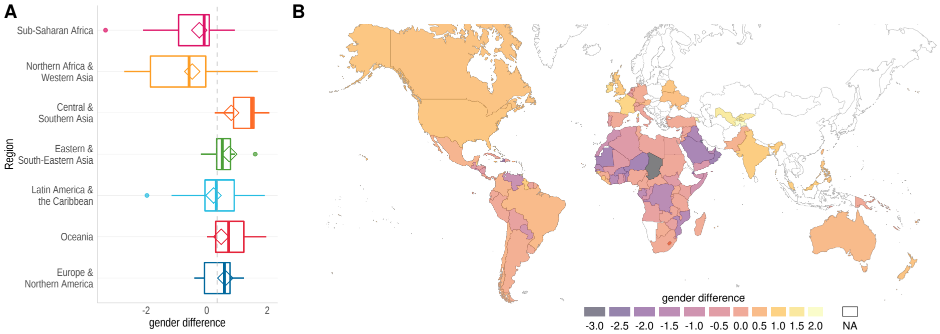

Figure S5: Gender differences in online language literacy across the world. (A) Summary of the standardized female-male differences across the seven geographical groups. Dashed line marks the global average (0.345, population-weighted), and a diamond indicates the population-weighted mean of the group. (B) Map of female-male differences available in our dataset.

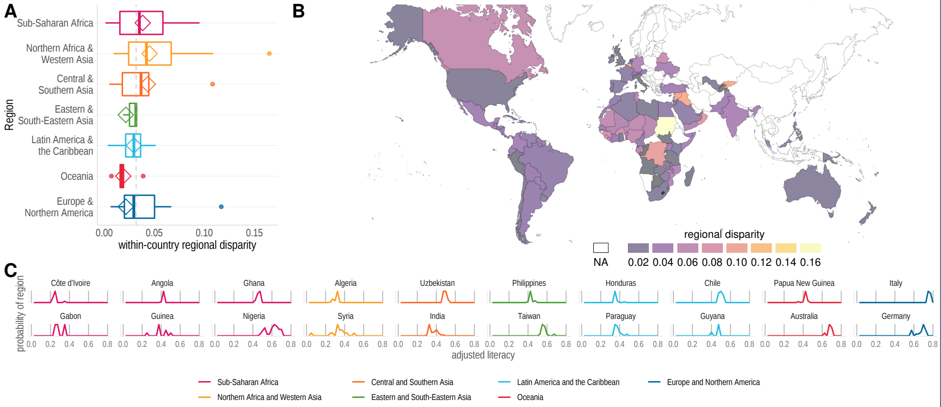

Figure S6: Within-country regional disparity in online language literacy across the world. (A) Summary of the within-country regional disparity across the seven geographical groups. Dashed line marks the global average (0.032, population-weighted), and a diamond indicates the population-weighted mean of the group. (B) Map of country-level regional disparity available in our dataset. (C) Countries with a higher or lower level of regional disparity from each geographical group.

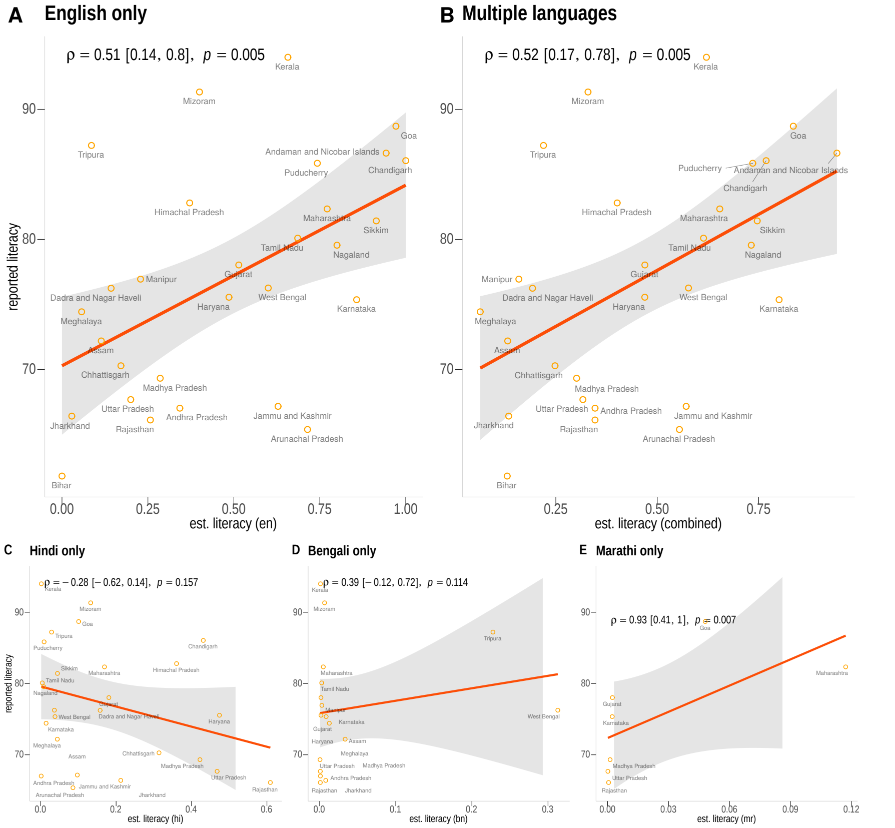

Figure S7: A multilingual country case study with India's user population. We include the six most used languages in India (English, Hindi, Bengali, Marathi, Telugu, and Tamil) to compare the literacy estimation based on the single representative language (English) with the estimation based on multiple languages. The results suggest that literacy estimation based on multiple languages has neglectable improvement over English-based estimation in terms of Spearman's rank correlations. The y-axis indicates the officially reported literacy level, and the x-axes indicate (A) the language literacy estimated using the regions' public posts in English, (B) the estimation using multiple languages combined, and (C-E) the estimation using posts in Hindi, Bengali, and Marathi only, respectively.

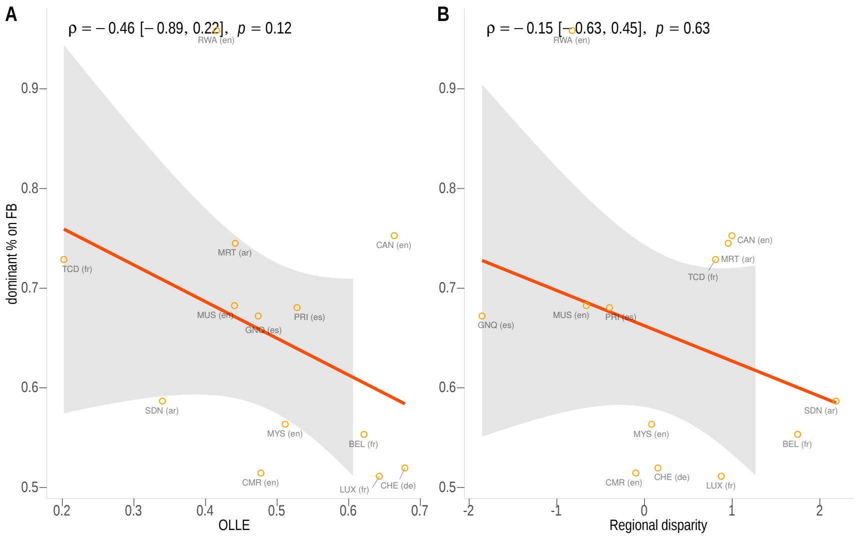

Figure S8: The relationship between the user percentage of the dominant language (y-axis) and (A) `OLLE`, or (B) regional disparity, in 13 multilingual countries studied. We do not observe a clear systematic bias. This serves as a robust check to see whether there is a potential bias in the set of multilingual countries.

Table S1: Knee points detected based on the Facebook popularity curves shown in Fig. S1. The two knee points, measured in 1,000 words, determine the LoFF words in each language. For example, the LoFF words in English correspond to the words in the fastText '`en`' dictionary that are ranked between 5,000 to 9,000 in the decreasing order of word frequency.

Language	en	es	fr	ar	de	zh	pt	it	ru	tr	ml	ms
k_0	5	5	5	5	5	5	5	5	5	5	5	6
k_1	9	9	9	9	9	16	9	11	8	11	10	21

Table S2: OLS for predicting the reported literacy rate with online literacy estimates. Model (a) is the fixed-effect model accounting for language-specific bias. The calibrated online literacy estimates (`OLLE`'s) are produced using model (a). The observations include all countries having online literacy estimates and corresponding predictors. Countries without sufficient Internet penetration ($< 25\%$) are excluded to obtain reliable calibrated models. For comparison, models (b,c) include additional predictors, the Internet penetration and income. All variables were transformed to better fit a normal distribution.

	DV: reported literacy rate		
	(a)	(b)	(c)
est. literacy	0.80^{***} (0.61, 0.99)	0.52^{***} (0.27, 0.76)	0.53^{***} (0.27, 0.79)
% Internet		0.32^{***} (0.13, 0.50)	0.49^{***} (0.19, 0.80)
income			−0.19 (−0.48, 0.11)
language [de]	1.43^{***} (0.72, 2.14)	1.22^{***} (0.53, 1.90)	1.29^{***} (0.59, 1.99)
language [en]	0.94^{***} (0.55, 1.33)	0.98^{***} (0.61, 1.35)	0.97^{***} (0.58, 1.37)
language [es]	1.31^{***} (0.82, 1.80)	1.15^{***} (0.68, 1.62)	1.25^{***} (0.76, 1.74)
language [fr]	1.43^{***} (0.79, 2.06)	1.13^{***} (0.51, 1.75)	1.30^{***} (0.57, 2.03)
language [it]	1.24^{**} (0.26, 2.22)	1.75^{***} (0.78, 2.73)	1.57^{**} (0.25, 2.90)
language [ms]	−0.75 (−1.75, 0.25)	−0.55 (−1.50, 0.40)	−0.52 (−1.48, 0.44)
language [nl]	0.91^{**} (0.21, 1.62)	0.78^{**} (0.10, 1.45)	0.97^{**} (0.21, 1.73)
language [pt]	0.62^{*} (−0.06, 1.29)	0.61^{*} (−0.03, 1.26)	0.66^{*} (0.01, 1.31)
language [ru]	3.41^{***} (2.64, 4.17)	3.14^{***} (2.41, 3.88)	3.10^{***} (2.34, 3.86)
language [tr]	0.86^{*} (−0.09, 1.81)	0.96^{**} (0.05, 1.86)	1.04^{**} (0.12, 1.95)
language [zh]	−0.62 (−1.64, 0.40)	−0.35 (−1.32, 0.63)	
Constant	-0.94^{***} (−1.27, −0.61)	-0.90^{***} (−1.21, −0.59)	-0.94^{***} (−1.26, −0.61)
OOS correlation ρ	0.78	0.8	0.79
OOS RMSE	0.7	0.66	0.68
OOS R^2	0.51	0.57	0.55
Observations	98	98	86
R^2	0.64	0.68	0.69
Adjusted R^2	0.59	0.63	0.64
AIC	205.52	195.32	172.75
BIC	241.70	234.10	209.57
Residual Std. Error	0.64 (df = 85)	0.61 (df = 84)	0.61 (df = 72)
F Statistic	12.49^{***} (df = 12; 85)	13.76^{***} (df = 13; 84)	12.52^{***} (df = 13; 72)

Note: *p<0.1; **p<0.05; ***p<0.01

The out-of-sample (OOS) Spearman correlation ρ, RMSE, and R^2 are obtained using leave-one-out cross-validation.

Table S3: OLS for predicting the reported literacy rate without online literacy estimates. All variables were transformed to better fit a normal distribution.

	DV: reported literacy rate		
	(a)	(b)	(c)
% Internet	0.54^{***} (0.37, 0.71)	0.58^{***} (0.43, 0.73)	0.74^{***} (0.43, 1.05)
income			-0.15 (-0.47, 0.18)
language [de]		1.14^{***} (0.40, 1.89)	1.22^{***} (0.45, 1.98)
language [en]		0.96^{***} (0.55, 1.36)	0.95^{***} (0.52, 1.38)
language [es]		0.69^{***} (0.23, 1.14)	0.79^{***} (0.31, 1.27)
language [fr]		0.64^{**} (0.02, 1.27)	0.75^{*} (0.004, 1.50)
language [it]		2.55^{***} (1.57, 3.53)	2.34^{***} (0.95, 3.74)
language [ms]		0.08 (-0.90, 1.07)	0.14 (-0.86, 1.14)
language [nl]		0.70^{*} (-0.04, 1.43)	0.95^{**} (0.11, 1.78)
language [pt]		0.33 (-0.35, 1.01)	0.40 (-0.30, 1.10)
language [ru]		2.51^{***} (1.78, 3.25)	2.49^{***} (1.72, 3.26)
language [tr]		1.17^{**} (0.20, 2.15)	1.27^{**} (0.27, 2.26)
language [zh]		0.42 (-0.56, 1.41)	
Constant	0.00 (-0.17, 0.17)	-0.78^{***} (-1.11, -0.44)	-0.83^{***} (-1.18, -0.48)
OOS correlation ρ	0.52	0.73	0.72
OOS RMSE	0.86	0.7	0.73
OOS R^2	0.26	0.51	0.48
Observations	98	98	86
R^2	0.29	0.62	0.62
Adjusted R^2	0.28	0.56	0.56
AIC	249.67	211.51	188.23
BIC	257.42	247.70	222.59
Residual Std. Error	0.85 (df = 96)	0.66 (df = 85)	0.67 (df = 73)
F Statistic	39.04^{***} (df = 1; 96)	11.32^{***} (df = 12; 85)	10.10^{***} (df = 12; 73)

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

The out-of-sample (OOS) Spearman correlation ρ, RMSE, and R^2 are obtained using leave-one-out cross-validation.

Table S4: Online language literacy estimates for all countries included in this study. Measures include: average relative LoFF words, OLLE ($N = 167$), female-male gender gap ($N = 160$) and regional disparity ($N = 119$).

country	code	big word ($\bar{w}$)	OLLE	female-male gap	regional disparity	no. regions	rep. language
Algeria	DZA	0.584	0.309	-0.584	0.022	48	ar
American Samoa	ASM	0.811	0.593	1.744	–	–	en
Angola	AGO	0.634	0.419	-0.281	0.008	17	pt
Anguilla	AIA	0.521	0.416	0.372	–	–	en
Antigua & Barbuda	ATG	0.715	0.519	1.422	–	–	en
Argentina	ARG	0.584	0.505	0.166	0.03	24	es
Armenia	ARM	0.308	0.624	1.684	0.047	10	ru
Aruba	ABW	0.785	0.58	0.476	–	–	nl
Australia	AUS	1.004	0.688	0.456	0.019	8	en
Austria	AUT	0.786	0.667	0.481	0.03	9	de
Bahamas	BHS	0.687	0.501	1.056	0.026	2	en
Bahrain	BHR	0.784	0.429	-2.293	0.084	2	ar
Barbados	BRB	0.743	0.544	1.325	–	–	en
Belarus	BLR	0.709	0.868	0.746	0.067	7	ru
Belgium	BEL	0.876	0.621	-0.411	0.117	3	nl
Belize	BLZ	0.729	0.528	1.011	0.021	4	en
Benin	BEN	0.548	0.498	-1.861	0.042	12	fr
Bermuda	BMU	0.955	0.659	0.551	–	–	en
Bolivia	BOL	0.437	0.439	-0.262	0.033	9	es
Botswana	BWA	0.535	0.422	0.144	0.004	8	en
Brazil	BRA	0.557	0.376	0.469	0.026	27	pt
British Virgin Islands	VGB	0.688	0.501	1.408	–	–	en
Brunei	BRN	1.522	0.468	0.465	–	–	ms
Burkina Faso	BFA	0.303	0.327	-1.921	0.074	6	fr
Burundi	BDI	0.141	0.211	-0.324	–	–	fr
Cameroon	CMR	0.638	0.477	0.016	0.027	10	en
Canada	CAN	0.965	0.664	0.695	0.065	13	en
Cape Verde	CPV	0.682	0.44	0.379	0.029	6	pt
Caribbean Netherlands	BES	0.849	0.609	–	–	–	nl
Cayman Islands	CYM	0.872	0.618	0.949	–	–	en
Central African Republic	CAF	0.317	0.36	0	–	–	fr
Chad	TCD	0.418	0.202	-3.344	0.056	2	ar
Chile	CHL	0.585	0.512	0.37	0.015	15	es
Colombia	COL	0.436	0.434	0.328	0.039	32	es
Comoros	COM	0.238	0.027	–	–	–	ar
Congo - Brazzaville	COG	0.316	0.36	-0.791	0.001	2	fr
Congo - Kinshasa	COD	0.312	0.359	-1.552	0.095	11	fr
Cook Islands	COK	0.848	0.61	–	–	–	en
Costa Rica	CRI	0.486	0.462	-0.039	0.043	7	es
Côte d’Ivoire	CIV	0.299	0.313	-1.604	0.032	14	fr
Cuba	CUB	0.475	0.459	-0.078	0.018	16	es
Curaçao	CUW	0.795	0.587	0.022	–	–	nl
Cyprus	CYP	0.846	0.601	-0.159	0.009	3	tr
Djibouti	DJI	0.468	0.258	-2.006	–	–	ar
Dominica	DMA	0.621	0.467	0.227	–	–	en
Dominican Republic	DOM	0.386	0.352	-0.189	0.029	31	es
Ecuador	ECU	0.451	0.453	0.002	0.034	24	es
Egypt	EGY	0.68	0.347	-0.145	0.028	27	ar
El Salvador	SLV	0.435	0.429	-0.937	0.033	14	es
Equatorial Guinea	GNQ	0.536	0.474	0.119	0.007	2	es
Fiji	FJI	0.678	0.492	0.482	0.016	3	en
France	FRA	0.767	0.632	1.094	0.025	22	fr
French Guiana	GUF	0.473	0.479	0.105	–	–	fr
French Polynesia	PYF	0.494	0.484	0.941	–	–	fr
Gabon	GAB	0.311	0.359	-1.445	0.059	3	fr
Gambia	GMB	0.522	0.416	-0.048	0.089	2	en
Germany	DEU	0.854	0.695	0.113	0.05	16	de
Ghana	GHA	0.576	0.448	-0.697	0.014	10	en
Gibraltar	GIB	0.945	0.651	0.774	–	–	en
Grenada	GRD	0.704	0.507	1.918	0.051	2	en
Guadeloupe	GLP	0.613	0.538	0.882	–	–	fr
Guam	GUM	0.772	0.567	1.475	–	–	en
Guatemala	GTM	0.403	0.383	-0.526	0.039	22	es
Guinea	GIN	0.348	0.366	0.085	0.076	8	fr
Guinea-Bissau	GNB	0.299	0.183	0.842	–	–	pt
Guyana	GUY	0.65	0.48	0.871	0.035	4	en
Haiti	HTI	0.28	0.265	-1.976	0.047	9	fr
Honduras	HND	0.387	0.364	-0.281	0.024	18	es
Hong Kong SAR China	HKG	2.021	0.57	0.02	–	–	zh
India	IND	0.423	0.361	0.887	0.044	34	en
Iraq	IRQ	0.745	0.399	-1.859	0.109	19	ar
Ireland	IRL	0.944	0.65	1.235	0.02	26	en
Isle of Man	IMN	1.086	0.693	0.335	–	–	en
Italy	ITA	1.159	0.757	-0.109	0.006	20	it
Jamaica	JAM	0.561	0.437	0.772	0.021	11	en
Jersey	JEY	0.949	0.655	1.234	–	–	en
Jordan	JOR	0.918	0.496	0.01	0.02	12	ar
Kenya	KEN	0.62	0.467	-0.208	0.021	8	en
Kiribati	KIR	0.557	0.433	0.751	–	–	en
Kuwait	KWT	0.759	0.405	-2.727	0.061	6	ar
Kyrgyzstan	KGZ	0.299	0.599	1.485	0.108	2	ru
Lebanon	LBN	0.734	0.386	-0.771	0.052	6	ar
Lesotho	LSO	0.485	0.41	-0.03	0.088	6	en
Liberia	LBR	0.74	0.538	0.483	0.053	2	en
Libya	LBY	0.697	0.355	-0.143	0.017	19	ar
Liechtenstein	LIE	0.779	0.656	–	–	–	de
Luxembourg	LUX	0.775	0.643	1.075	0.059	2	fr
Macau SAR China	MAC	1.69	0.509	0.433	–	–	zh
Madagascar	MDG	0.252	0.253	0.078	0.012	18	fr
Malawi	MWI	0.522	0.416	-0.09	0.036	5	en
Malaysia	MYS	1.807	0.511	0.936	0.032	15	ms
Mali	MLI	0.244	0.25	-0.537	0.051	6	fr
Malta	MLT	0.718	0.525	0.733	0.029	4	en
Marshall Islands	MHL	0.729	0.531	1.173	–	–	en
Mauritania	MRT	0.8	0.442	-2.055	0.063	2	ar
Mauritius	MUS	0.568	0.44	0.59	0.018	9	en
Mayotte	MYT	0.406	0.406	0.773	–	–	fr
Mexico	MEX	0.428	0.418	0.318	0.037	32	es

Micronesia (Federated States of)	FSM	0.532	0.418	0.139	0.007	2	en
Monaco	MCO	0.883	0.7	–	–	–	fr
Morocco	MAR	0.628	0.322	-1.058	0.026	16	ar
Mozambique	MOZ	0.288	0.148	-1.678	0.051	10	pt
Nauru	NRU	0.611	0.465	–	–	–	en
Netherlands	NLD	0.839	0.603	-0.215	0.025	12	nl
New Caledonia	NCL	0.469	0.478	0.032	0.039	2	fr
New Zealand	NZL	1.001	0.679	0.686	0.016	16	en
Nicaragua	NIC	0.42	0.406	-0.553	0.031	17	es
Niger	NER	0.294	0.299	-2.103	0.061	4	fr
Nigeria	NGA	0.814	0.597	0.002	0.064	37	en
Northern Mariana Islands	MNP	0.708	0.51	1.28	–	–	en
Oman	OMN	0.796	0.441	-2.218	0.088	3	ar
Pakistan	PAK	0.455	0.403	0.262	0.037	8	en
Palestinian Territories	PSE	0.905	0.49	0.266	0.019	2	ar
Panama	PAN	0.446	0.449	0.005	0.043	8	es
Papua New Guinea	PNG	0.539	0.423	0.013	0.019	17	en
Paraguay	PRY	0.429	0.424	-0.883	0.035	17	es
Peru	PER	0.535	0.47	0.001	0.009	25	es
Philippines	PHL	0.551	0.426	0.735	0.019	17	en
Portugal	PRT	0.711	0.458	-0.41	0.018	20	pt
Puerto Rico	PRI	0.632	0.528	0.375	0.022	68	es
Qatar	QAT	0.812	0.444	-2.442	–	–	ar
Réunion	REU	0.579	0.522	0.461	–	–	fr
Rwanda	RWA	0.521	0.416	-0.037	0.015	5	en
Saint Martin (French part)	MAF	0.699	0.576	–	–	–	fr
Samoa	WSM	0.655	0.483	0.725	–	–	en
San Marino	SMR	1.208	0.768	-0.191	–	–	it
São Tomé & Príncipe	STP	0.399	0.265	0.168	–	–	pt
Saudi Arabia	SAU	0.764	0.409	-1.981	0.026	13	ar
Senegal	SEN	0.285	0.267	-1.618	0.035	14	fr
Seychelles	SYC	0.66	0.486	0.931	–	–	en
Sierra Leone	SLE	0.759	0.553	0.714	0.076	4	en
Singapore	SGP	0.742	0.541	1.602	–	–	en
Solomon Islands	SLB	0.545	0.424	0.081	–	–	en
Somalia	SOM	0.608	0.317	-0.959	0.04	5	ar
South Africa	ZAF	0.514	0.414	0.052	0.018	9	en
South Sudan	SSD	0.605	0.464	-0.514	0.01	2	en
Spain	ESP	0.892	0.684	0.033	0.023	17	es
St. Kitts & Nevis	KNA	0.753	0.55	1.657	0.003	2	en
St. Lucia	LCA	0.631	0.473	0.792	0.02	3	en
St. Vincent & Grenadines	VCT	0.641	0.478	0.183	0.043	2	en
Sudan	SDN	0.673	0.34	-0.032	0.165	9	ar
Suriname	SUR	0.572	0.442	-0.185	0.031	2	nl
Swaziland	SWZ	0.555	0.43	-0.326	0.016	4	en
Switzerland	CHE	0.812	0.678	-0.113	0.033	22	de
Syria	SYR	0.7	0.356	-0.107	0.105	14	ar
Taiwan	TWN	2.101	0.583	0.565	0.033	16	zh
Tajikistan	TJK	0.229	0.527	2.07	0.005	2	ru
Tanzania	TZA	0.444	0.397	-0.037	0.012	19	en
Timor-Leste	TLS	0.416	0.29	-0.193	–	–	pt
Togo	TGO	0.273	0.262	-0.839	–	–	fr
Tonga	TON	0.739	0.538	1.969	–	–	en
Trinidad & Tobago	TTO	0.677	0.489	0.948	0.028	13	en
Tunisia	TUN	0.709	0.362	0.058	0.05	24	ar
Turkey	TUR	0.835	0.593	0.055	0.037	78	tr
Turks & Caicos Islands	TCA	0.676	0.489	0.935	–	–	en
Uganda	UGA	0.624	0.469	0.06	0.035	22	en
Ukraine	UKR	0.584	0.806	0.625	0.044	27	ru
United Arab Emirates	ARE	0.718	0.374	-1.421	0.025	6	ar
United Kingdom	GBR	0.987	0.671	0.807	0.011	4	en
United States	USA	0.951	0.657	0.755	0.017	51	en
Uruguay	URY	0.572	0.499	0.104	0.02	19	es
U.S. Virgin Islands	VIR	0.842	0.607	1.41	–	–	en
Uzbekistan	UZB	0.223	0.524	1.558	0.018	9	ru
Venezuela	VEN	0.437	0.444	-1.165	0.028	24	es
Yemen	YEM	0.979	0.52	-0.703	0.048	7	ar
Zambia	ZMB	0.568	0.441	-0.083	0.016	9	en
Zimbabwe	ZWE	0.602	0.464	-0.295	0.024	8	en

The representative language for each country is chosen as the most used language by the country's population observed on Facebook.

Table S5: Variables related to gender gap analysis. Reported N is the number of countries matched with our data.

Statistic	N	Mean	St. Dev.	Min	Max	Median	Definition	Source
OLLE gap	160	0.053	1.002	−3.344	2.070	0.080	female-map gap in OLLE	
offline literacy (all)	143	84.181	19.263	19.100	100.000	93.464	literacy rate	UNESCO [?]
offline literacy (gap)	114	−0.068	0.090	−0.301	0.182	−0.036	female-male gap in literacy	UNESCO [?]
education (all)	106	7.884	2.741	1.880	13.180	8.085	mean schooling years	Barro-Lee Educational Attainment Data [?]
education (gap)	106	−0.517	0.968	−3.250	1.600	−0.420	female-map gap in schooling years	Barro-Lee Educational Attainment Data [?]
% Internet (all)	151	0.520	0.284	0.020	0.984	0.555	overall Internet penetration	ITU Internet gender gap [?]
% Internet (gap)	128	0.880	0.123	0.545	1.000	0.919	femal-male gap in Internet penetration	Digital gender gap (U. Oxford) [?]
civic (all)	123	0.695	0.211	0.105	0.973	0.746	overall civic society participation	V-Dem Institute [57]
civic (women)	123	0.720	0.173	0.234	0.937	0.775	women's civic society participation	V-Dem Institute [57]
GII	107	0.398	0.194	0.040	0.835	0.424	Gender Inequality Index	HDRO [?]

Table S6: Correlations among variables related to gender gap or gender equity. All correlations are reported using Spearman rank correlation coefficients.

	OLLE gap	OLLE	offline literacy (all)	offline literacy (gap)	eduation (all)	education (gap)	% Internet (all)	% Internet (gap)	civic (all)	civic (women)
OLLE	0.585***									
off. literacy (all)	0.524***	0.740***								
off. literacy (gap)	0.420***	0.438***	0.709***							
eduation (all)	0.587***	0.739***	0.872***	0.694***						
education (gap)	0.225*	0.207*	0.515***	0.787***	0.455***					
% Internet (all)	0.304***	0.573***	0.748***	0.582***	0.743***	0.434***				
% Internet (gap)	0.487***	0.533***	0.663***	0.779***	0.730***	0.620***	0.617***			
civic (all)	0.287**	0.355***	0.159	-0.019	0.358***	0.028	0.228*	0.328***		
civic (women)	0.479***	0.484***	0.394***	0.317**	0.564***	0.252*	0.429***	0.598***	0.695***	
GII	-0.391***	-0.624***	-0.847***	-0.610***	-0.838***	-0.475***	-0.872***	-0.647***	-0.219*	-0.414***

significance levels: $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table S7: Variables related to country resource access and inequality. Reported N is the number of countries matched with our data.

Statistic	N	Mean	St. Dev.	Min	Max	Median	Definition	Source
regional disparity	119	0.037	0.027	0.001	0.165	0.030	St. Dev. of sub-national OLLE	
income	115	16.193	17.054	0.800	71.160	9.359	GNI per capita in 1,000 US$ (2011 PPP)	HDI [?]
Gini index	94	40.004	7.973	25.000	63.000	40.200	Gini coefficient for income	HDR [?]
education (all)	95	7.864	2.753	1.880	13.180	7.970	mean schooling years	Barro-Lee Educational Attainment Data [?]
unequal education	102	21.025	14.073	0.800	49.300	17.450	Inequality in education	HDR [?]
% Internet (all)	119	0.507	0.275	0.020	0.980	0.508	overall Internet penetration	ITU Internet gender gap [?]
civic (all)	110	0.704	0.213	0.105	0.973	0.764	overall civic society participation	V-Dem Institute [57]

Table S8: Correlations among variables related to country resource access and inequality. All correlations are reported using Spearman rank correlation coefficients.

	regional disparity	OLLE	income	Gini index	education (all)	unequal education	% Internet (all)
regional disparity							
OLLE	-0.158						
income	-0.156	0.476***					
Gini index	-0.271**	-0.346***	-0.260*				
eduation (all)	-0.312**	0.706***	0.745***	-0.303**			
unequal education	0.283**	-0.681***	-0.703***	0.240*	-0.875***		
% Internet (all)	-0.043	0.532***	0.905***	-0.306**	0.751***	-0.735***	
civic (all)	-0.012	0.344***	0.192*	-0.154	0.366***	-0.207*	0.237*

significance levels: $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table S9: OLS for predicting female-male gender gap in online literacy estimates. Models (d) corresponds to the figure in the main text. Models (b,c,e,f) include alternative interaction terms. Values in parentheses are the lower and upper bounds of the 95% confidence intervals of the estimated effects.

	DV: female-male gender gap					
	(a)	(b)	(c)	(d)	(e)	(f)
women civic	0.32*** (0.15, 0.48)	0.33*** (0.14, 0.52)	0.35*** (0.17, 0.52)	0.38*** (0.20, 0.56)	0.37*** (0.18, 0.56)	0.40*** (0.22, 0.57)
(women civic):(% Internet)	0.25*** (0.10, 0.41)			0.13* (−0.02, 0.29)		
(women civic):(education)		0.28 (−0.10, 0.66)			0.19 (−0.14, 0.53)	
education (all)	0.25*** (0.08, 0.42)	0.37*** (0.13, 0.61)	0.90** (0.19, 1.61)	0.20** (0.05, 0.35)	0.28** (0.06, 0.49)	0.66** (0.03, 1.28)
(education):(% Internet)			0.51* (−0.02, 1.05)			0.36 (−0.11, 0.84)
% Internet (all)	−0.02 (−0.21, 0.16)	−0.08 (−0.27, 0.11)	−0.23* (−0.50, 0.03)	−0.22* (−0.45, 0.01)	−0.28** (−0.50, −0.06)	−0.41*** (−0.69, −0.13)
Central/Southern/Eastern Asia				1.66*** (1.05, 2.27)	1.77*** (1.17, 2.37)	1.80*** (1.22, 2.39)
Europe/Oceania/Northern America				0.63** (0.03, 1.22)	0.76*** (0.20, 1.33)	0.79*** (0.25, 1.34)
Latin America & the Caribbean				0.29 (−0.15, 0.74)	0.32 (−0.12, 0.77)	0.38* (−0.06, 0.82)
Northern Africa & Western Asia				0.29 (−0.26, 0.84)	0.30 (−0.26, 0.85)	0.41 (−0.14, 0.96)
Constant	−0.35*** (−0.51, −0.18)	−0.30*** (−0.48, −0.13)	−0.47*** (−0.75, −0.19)	−0.69*** (−1.02, −0.37)	−0.72*** (−1.05, −0.40)	−0.89*** (−1.28, −0.50)
Out-of-sample RMSE	0.8	0.83	0.83	0.72	0.73	0.73
Out-of-sample R2	0.24	0.21	0.2	0.39	0.38	0.38
Observations	101	101	101	101	101	101
R^2	0.34	0.29	0.30	0.51	0.51	0.51
Adjusted R^2	0.31	0.26	0.27	0.47	0.46	0.47
AIC	240.92	248.67	247.25	217.60	219.21	218.11
BIC	256.61	264.36	262.94	243.75	245.36	244.26
Residual Std. Error	0.77 (df = 96)	0.80 (df = 96)	0.80 (df = 96)	0.67 (df = 92)	0.68 (df = 92)	0.68 (df = 92)
F Statistic	12.26*** (df = 4; 96)	9.58*** (df = 4; 96)	10.06*** (df = 4; 96)	12.19*** (df = 8; 92)	11.82*** (df = 8; 92)	12.07*** (df = 8; 92)
Note:						*p<0.1; **p<0.05; ***p<0.01

Table S10: OLS for predicting female-male gender gap in online literacy estimates. Models include alternative predictors and no interaction term. Values in parentheses are the lower and upper bounds of the 95% confidence intervals of the estimated effects.

	DV: female-male gender gap					
	(a)	(b)	(c)	(d)	(e)	(f)
women civic	0.39*** (0.22, 0.56)	0.39*** (0.22, 0.57)	0.40*** (0.20, 0.60)	0.37*** (0.15, 0.59)		
civic (all)					0.21** (0.05, 0.38)	0.21** (0.01, 0.41)
education (gap)		0.05 (−0.12, 0.23)		0.02 (−0.19, 0.24)		
education (all)	0.24*** (0.07, 0.42)		0.24** (0.06, 0.43)		0.27*** (0.09, 0.46)	0.24** (0.04, 0.43)
% Internet (gap)			−0.03 (−0.25, 0.20)	0.10 (−0.17, 0.37)		0.15 (−0.06, 0.36)
% Internet (all)	−0.05 (−0.24, 0.14)	0.04 (−0.16, 0.23)			0.07 (−0.12, 0.26)	
Constant	−0.25*** (−0.40, −0.09)	−0.25*** (−0.41, −0.09)	−0.26*** (−0.42, −0.09)	−0.25*** (−0.42, −0.08)	−0.24*** (−0.41, −0.07)	−0.25*** (−0.42, −0.07)
Out-of-sample RMSE	0.83	0.87	0.85	0.88	0.91	0.92
Out-of-sample R2	0.19	0.13	0.21	0.13	0.09	0.1
Observations	101	101	98	98	101	98
R^2	0.27	0.22	0.28	0.22	0.17	0.19
Adjusted R^2	0.25	0.19	0.25	0.20	0.15	0.17
AIC	248.88	256.07	243.76	250.52	261.40	254.21
BIC	261.96	269.14	256.69	263.44	274.48	267.13
Residual Std. Error	0.81 (df = 97)	0.83 (df = 97)	0.81 (df = 94)	0.84 (df = 94)	0.86 (df = 97)	0.86 (df = 94)
F Statistic	11.93*** (df = 3; 97)	8.89*** (df = 3; 97)	11.96*** (df = 3; 94)	9.07*** (df = 3; 94)	6.77*** (df = 3; 97)	7.58*** (df = 3; 94)
Note:						*p<0.1; **p<0.05; ***p<0.01 Models (c,d,f) have fewer observations (and slightly lower prediction error) due to missing data in the new predictor.

Table S11: OLS for predicting female-male gender gap in online literacy estimates. Models include alternative predictors and controls for geographical groups. Values in parentheses are the lower and upper bounds of the 95% confidence intervals of the estimated effects.

	DV: female-male gender gap					
	(a)	(b)	(c)	(d)	(e)	(f)
women civic	0.41^{***} (0.23, 0.58)	0.43^{***} (0.26, 0.61)	0.35^{***} (0.16, 0.53)	0.37^{***} (0.17, 0.56)		
civic (all)					0.16^{*} (-0.02, 0.34)	0.16 (-0.04, 0.37)
education (gap)		0.13^{*} (-0.02, 0.29)		0.10 (-0.08, 0.29)		
education (all)	0.19^{**} (0.04, 0.34)		0.19^{**} (0.02, 0.36)		0.21^{**} (0.05, 0.38)	0.18^{**} (0.01, 0.36)
% Internet (gap)			-0.07 (-0.30, 0.15)	-0.07 (-0.33, 0.20)		0.08 (-0.14, 0.30)
% Internet (all)	-0.28^{**} (-0.50, -0.06)	-0.30^{**} (-0.54, -0.06)			-0.11 (-0.34, 0.11)	
Central/Southern/Eastern Asia	1.81^{***} (1.22, 2.41)	1.99^{***} (1.40, 2.59)	1.48^{***} (0.92, 2.03)	1.64^{***} (1.09, 2.19)	1.67^{***} (1.02, 2.32)	1.50^{***} (0.91, 2.10)
Europe/Oceania/Northern America	0.84^{***} (0.30, 1.39)	0.97^{***} (0.42, 1.53)	0.46^{*} (-0.03, 0.96)	0.58^{**} (0.07, 1.08)	0.83^{***} (0.24, 1.43)	0.53^{*} (0.002, 1.05)
Latin America & the Caribbean	0.37 (-0.07, 0.81)	0.44^{*} (-0.004, 0.88)	0.18 (-0.28, 0.65)	0.25 (-0.22, 0.72)	0.39 (-0.09, 0.88)	0.17 (-0.32, 0.67)
Northern Africa & Western Asia	0.35 (-0.20, 0.90)	0.48^{*} (-0.07, 1.04)	-0.16 (-0.63, 0.31)	-0.04 (-0.50, 0.42)	0.03 (-0.58, 0.65)	-0.20 (-0.74, 0.33)
Constant	-0.73^{***} (-1.06, -0.40)	-0.81^{***} (-1.14, -0.48)	-0.50^{***} (-0.80, -0.20)	-0.58^{***} (-0.88, -0.28)	-0.65^{***} (-1.02, -0.28)	-0.50^{***} (-0.82, -0.17)
Out-of-sample RMSE	0.73	0.75	0.76	0.8	0.8	0.82
Out-of-sample R^2	0.37	0.35	0.35	0.31	0.28	0.29
Observations	101	101	98	98	101	98
R^2	0.50	0.48	0.48	0.46	0.41	0.42
Adjusted R^2	0.46	0.44	0.44	0.42	0.36	0.37
AIC	218.59	222.02	218.82	222.81	235.76	230.10
BIC	242.12	245.56	242.09	246.07	259.30	253.36
Residual Std. Error	0.68 (df = 93)	0.69 (df = 93)	0.70 (df = 90)	0.72 (df = 90)	0.74 (df = 93)	0.75 (df = 90)
F Statistic	13.29^{***} (df = 7; 93)	12.40^{***} (df = 7; 93)	12.00^{***} (df = 7; 90)	11.01^{***} (df = 7; 90)	9.13^{***} (df = 7; 93)	9.30^{***} (df = 7; 90)

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Models (c,d,f) have fewer observations (and slightly lower prediction error) due to missing data in the new predictor.

Table S12: OLS for predicting within-country regional disparity in online literacy estimates. Models (d) corresponds to the figure in the main text. Models (b,c,e,f) include alternative interaction terms. Values in parentheses are the lower and upper bounds of the 95% confidence intervals of the estimated effects.

	DV: within-country regional disparity					
	(a)	(b)	(c)	(d)	(e)	(f)
unequal edu	0.47** (0.10, 0.84)	0.34** (0.02, 0.66)	0.32* (−0.02, 0.66)	0.41** (0.03, 0.80)	0.25 (−0.10, 0.60)	0.25 (−0.12, 0.63)
(unequal edu):(Gini)	0.18 (−0.07, 0.44)			0.25* (−0.02, 0.51)		
(unequal edu):(% Internet)		−0.09 (−0.36, 0.18)			−0.15 (−0.46, 0.17)	
(unequal edu):(education)			0.003 (−0.37, 0.38)			0.01 (−0.38, 0.41)
Gini	−0.22** (−0.42, −0.02)	−0.22** (−0.43, −0.02)	−0.24** (−0.46, −0.02)	−0.35** (−0.61, −0.09)	−0.32** (−0.61, −0.04)	−0.37** (−0.67, −0.08)
% Internet (all)	0.40** (0.08, 0.72)	0.33* (0.0000, 0.66)	0.35** (0.01, 0.70)	0.46** (0.08, 0.83)	0.34* (−0.05, 0.74)	0.41* (−0.02, 0.83)
education (all)	−0.29** (−0.55, −0.02)	−0.27* (−0.56, 0.02)	−0.31 (−0.83, 0.21)	−0.33** (−0.60, −0.06)	−0.29* (−0.59, −0.003)	−0.36 (−0.90, 0.19)
Central/Southern/Eastern Asia				−0.17 (−1.06, 0.72)	−0.05 (−1.04, 0.94)	−0.24 (−1.16, 0.68)
Europe/Oceania/Northern America				−0.54 (−1.51, 0.43)	−0.42 (−1.40, 0.56)	−0.47 (−1.47, 0.54)
Latin America & the Caribbean				0.26 (−0.36, 0.88)	0.27 (−0.38, 0.92)	0.19 (−0.46, 0.84)
Northern Africa & Western Asia				−0.09 (−0.89, 0.72)	−0.03 (−0.90, 0.83)	−0.18 (−1.03, 0.67)
Constant	−0.04 (−0.24, 0.16)	−0.06 (−0.33, 0.21)	0.004 (−0.30, 0.31)	0.03 (−0.47, 0.52)	−0.06 (−0.66, 0.54)	0.11 (−0.51, 0.73)
Out-of-sample RMSE	0.9	0.92	0.92	0.92	0.95	0.95
Out-of-sample R^2	0.14	0.1	0.11	0.09	0.08	0.06
Observations	79	79	79	79	79	79
R^2	0.25	0.23	0.23	0.29	0.26	0.25
Adjusted R^2	0.19	0.18	0.17	0.19	0.17	0.16
AIC	207.37	209.02	209.48	210.91	213.57	214.53
BIC	223.96	225.60	226.07	236.98	239.63	240.59
Residual Std. Error	0.86 (df = 73)	0.86 (df = 73)	0.87 (df = 73)	0.86 (df = 69)	0.87 (df = 69)	0.88 (df = 69)
F Statistic	4.75*** (df = 5; 73)	4.36*** (df = 5; 73)	4.24*** (df = 5; 73)	3.09*** (df = 9; 69)	2.73*** (df = 9; 69)	2.61** (df = 9; 69)
Note:						$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table S13: OLS for predicting within-country regional disparity in online literacy estimates. Models include alternative predictors and no interaction term. Values in parentheses are the lower and upper bounds of the 95% confidence intervals of the estimated effects.

	DV: within-country regional disparity					
	(a)	(b)	(c)	(d)	(e)	(f)
unequal edu	0.33^{**}	0.15	0.18	0.26	0.13	0.15
	(0.01, 0.64)	(−0.17, 0.46)	(−0.12, 0.47)	(−0.10, 0.61)	(−0.22, 0.48)	(−0.19, 0.50)
Gini	-0.24^{**}	-0.28^{***}	-0.28^{***}	-0.37^{***}	-0.34^{**}	-0.33^{**}
	(−0.44, −0.05)	(−0.49, −0.08)	(−0.48, −0.08)	(−0.63, −0.10)	(−0.61, −0.06)	(−0.60, −0.06)
% Internet (all)	0.35^{**}			0.40^{**}		
	(0.04, 0.67)			(0.02, 0.77)		
education (all)	-0.31^{**}	−0.20	−0.20	-0.34^{**}	−0.24	-0.27^{*}
	(−0.57, −0.05)	(−0.49, 0.08)	(−0.46, 0.05)	(−0.61, −0.06)	(−0.54, 0.07)	(−0.55, −0.003)
income		−0.01			−0.06	
		(−0.32, 0.31)			(−0.45, 0.33)	
civic (all)			0.08			0.17
			(−0.15, 0.32)			(−0.11, 0.45)
Central/Southern/Eastern Asia				−0.23	−0.03	0.13
				(−1.14, 0.67)	(−0.96, 0.89)	(−0.83, 1.08)
Europe/Oceania/Northern America				−0.46	0.11	0.02
				(−1.44, 0.52)	(−0.94, 1.17)	(−0.86, 0.90)
Latin America & the Caribbean				0.19	0.47	0.51
				(−0.43, 0.82)	(−0.16, 1.10)	(−0.09, 1.12)
Northern Africa & Western Asia				−0.17	0.22	0.39
				(−0.99, 0.64)	(−0.60, 1.05)	(−0.45, 1.22)
Constant	0.002	0.001	−0.02	0.10	−0.17	−0.23
	(−0.19, 0.19)	(−0.20, 0.20)	(−0.22, 0.19)	(−0.40, 0.59)	(−0.68, 0.34)	(−0.70, 0.25)
Out-of-sample RMSE	0.91	0.92	0.92	0.95	0.96	0.95
Out-of-sample R^2	0.12	0.09	0.1	0.08	0.05	0.05
Observations	79	79	79	79	79	79
R^2	0.23	0.18	0.18	0.25	0.21	0.22
Adjusted R^2	0.18	0.13	0.14	0.17	0.12	0.13
AIC	207.48	212.39	211.87	212.54	217.18	215.68
BIC	221.70	226.61	226.09	236.23	240.87	239.37
Residual Std. Error	0.86 (df = 74)	0.89 (df = 74)	0.89 (df = 74)	0.87 (df = 70)	0.90 (df = 70)	0.89 (df = 70)
F Statistic	5.38^{***} (df = 4; 74)	3.94^{***} (df = 4; 74)	4.09^{***} (df = 4; 74)	2.97^{***} (df = 8; 70)	2.30^{**} (df = 8; 70)	2.52^{**} (df = 8; 70)
Note:						*p<0.1; **p<0.05; ***p<0.01

Author details

[1]School of Computing and Information, University of Pittsburgh, Pittsburgh, PA, USA. [2]AImpower.org, Mountain View, CA, USA. [3]Meta, Menlo Park, CA, USA.

References

1. Kirsch, I.S., Jungeblut, A., Jenkins, L., Kolstad, A.: Adult literacy in america: A first look at the findings of the national adult literacy survey (2002). NATIONAL CENTER FOR EDUCATION STATISTICS. Access on 09/23/2022 at `https://nces.ed.gov/pubs93/93275.pdf`
2. Kutner, M., Greenberg, E., Jin, Y., Boyle, B., Hsu, Y.-c., Dunleavy, E.: Literacy in Everyday Life: Results from the 2003 National Assessment of Adult Literacy. NCES 2007-490. Report, U.S. Department of Education. Washington, DC: National Center for Education Statistics (2007)
3. Schütz, G., Ursprung, H.W., Wößmann, L.: Education policy and equality of opportunity. Kyklos **61**(2), 279–308 (2008)
4. Kutner, M., Greenberg, E., Jin, Y., Boyle, B., Hsu, Y.-c., Dunleavy, E., White, S.: Literacy in everyday life: Results from the 2003 national assessment of adult literacy (2017). (Accessed on 09/23/2022)
5. Ferrer, A., Green, D.A., Riddell, W.C.: The effect of literacy on immigrant earnings. Journal of Human Resources **41**(2), 380–410 (2006)
6. Bonikowska, A., Green, D.A., Riddell, W.C.: Literacy and the Labour Market: Cognitive Skills and Immigrant Earnings. Statistics Canada, Ottawa (2008)
7. Schwerdt, G., Wiederhold, S., Murray, T.S.: Literacy and growth: New evidence from PIAAC. Retrieved from PIAAC Gateway website: `http://piaacgateway.com/`. (Accessed on 06/30/2020) (2020)
8. Dewalt, D., Berkman, N., Sheridan, S., Lohr, K., Pignone, M.: Literacy and health outcomes: A systematic review of the literature. Journal of general internal medicine **19**, 1228–39 (2005). doi:10.1111/j.1525–1497.2004.40153.x
9. Desjardins, R., Thorn, W., Schleicher, A., Quintini, G., Pellizzari, M., Kis, V., Chung, J.E.: OECD Skills Outlook 2013: First Results from the Survey of Adult Skills. `http://dx.doi.org/10.1787/9789264204256-en`. Accessed on 11/23/2020 (2013)
10. Gerger, C.: Social linguistics and literacies: Ideology in discourses. social linguistics and literacies: Ideology in discourses. Ilha do Desterro (2008)
11. UNESCO Institute for Statistics: Literacy Rates Continue to Rise from One Generation to the Next. Retrieved September 22, 2022 from `http://uis.unesco.org/sites/default/files/documents/fs45-literacy-rates-continue-rise-generation-to-next-en-2017_0.pdf` (2017)
12. Mundial, G.B., UNICEF, et al.: Education 2030: Incheon declaration and framework for action: towards inclusive and equitable quality education and lifelong learning for all (2016)
13. Bach, A.J., Wolfson, T., Crowell, J.K.: Poverty, literacy, and social transformation: An interdisciplinary exploration of the digital divide. Journal of Media Literacy Education **10**(1), 22–41 (2018)
14. Relations, M.I.: Meta Reports Third Quarter 2022 Results. `https://investor.fb.com/investor-news/press-release-details/2022/Meta-Reports-Third-Quarter-2022-Results/default.aspx` (2022)
15. Relations, M.I.: Meta Earnings Presentation Q3 2022. `https://s21.q4cdn.com/399680738/files/doc_financials/2022/q3/Q3-2022_Earnings-Presentation.pdf` (2022)
16. Rammstedt, B., Maehler, D.B.: Introduction: PIAAC and its Methodological Challenges. methods, data, analyses **8**(2), 12 (2016)
17. SDG17: United Nations Sustainable Development Goals. `https://sdgs.un.org/goals`. Accessed on 10/15/2020 (2015)
18. Montoya, S.: 50 Years of International Literacy Day: Time to Develop New Literacy Data — UNESCO UIS. `http://uis.unesco.org/en/blog/50-years-international-literacy-day-time-develop-new-literacy-data`. (Accessed on 06/30/2020) (2016)
19. UNESCO Institute for Statistics: UIS Statistics. Retrieved September 22, 2022 from `http://data.uis.unesco.org/index.aspx?queryid=3445#` (2022)
20. NCES: Program for the International Assessment of Adult Competencies (PIAAC). `https://nces.ed.gov/surveys/piaac/`. (Accessed on 09/23/2022) (2012)
21. Hargittai, E.: An update on survey measures of web-oriented digital literacy. Social science computer review **27**(1), 130–137 (2009)
22. DiMaggio, P., Hargittai, E., *et al.*: From the "digital divide" to "digital inequality": Studying internet use as penetration increases. Princeton: Center for Arts and Cultural Policy Studies, Woodrow Wilson School, Princeton University **4**(1), 4–2 (2001)
23. Sprague, K., Grijpink, F., Manyika, J., Moodley, L., Chappuis, B., Pattabiraman, K., Bughin, J.: Offline and falling behind: Barriers to Internet adoption. Retrieved September 22, 2022 from `https://www.mckinsey.com/~/media/mckinsey/dotcom/client_service/high%20tech/pdfs/offline_and_falling_behind_full_report.ashx` (2014)
24. Ortiz-Ospina, E.: The rise of social media - Our World in Data. `https://ourworldindata.org/rise-of-social-media`. (Accessed on 11/20/2021) (2019)
25. Erstad, O., Gilje, N., de Lange, T.: Re-mixing multimodal resources: Multiliteracies and digital production in norwegian media education. Learning, Media and Technology **32**, 183–198 (2007). doi:10.1080/17439880701343394
26. Alvermann, D.E.: Why bother theorizing adolescents' online literacies for classroom practice and research? Journal of Adolescent & Adult Literacy **52**(1), 8–19 (2008). Accessed 2022-10-01
27. Greenhow, C., Robelia, B.: Old communication, new literacies: Social network sites as social learning resources. J. Computer-Mediated Communication **14**, 1130–1161 (2009). doi:10.1111/j.1083-6101.2009.01484.x
28. Perkel, D.: Copy and paste literacy? literacy practices in the production of a myspace profile. Informal Learning and Digital Media **49** (2010). doi:10.1590/S0103-18132010000200011

29. Greenhow, C., Gleason, B.: Twitteracy: Tweeting as a new literacy practice. The Educational Forum **76**(4), 464–478 (2012). doi:10.1080/00131725.2012.709032
30. Davies, J.: Facework on facebook as a new literacy practice. Computers & Education **59**(1), 19–29 (2012). doi:10.1016/j.compedu.2011.11.007. CAL 2011
31. Kress, G.: Literacy in the new media age, 1–190 (2003). doi:10.4324/9780203299234
32. Clark, C., Dugdale, G.: People's Writing: Attitudes, behaviour and the role of technology. `https://files.eric.ed.gov/fulltext/ED510271.pdf`. (Accessed on 09/30/2022) (2009)
33. Sabaruddin: Facebook utilisation to enhance english writing skill. English Language Teaching **12**(8), 37–43 (2019)
34. Black, R.W.: Just don't call them cartoons: The new literacy spaces of anime, manga, and fanfiction. In: Coiro, J., Knobel, M., Lankshear, C., Leu, D.J. (eds.) Handbook of Research on New Literacies, pp. 583–610. Taylor & Francis, New York (2008)
35. Kojo, D.B., Agyekum, B.O., Arthur, B.: Exploring the effects of social media on the reading culture of students in tamale technical university. Journal of Education and Practice **9**(7), 47–56 (2018)
36. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science **359**(6380), 1146–1151 (2018). doi:10.1126/science.aap9559. https://www.science.org/doi/pdf/10.1126/science.aap9559
37. Edelson, L., Nguyen, M.-K., Goldstein, I., Goga, O., McCoy, D., Lauinger, T.: Understanding engagement with u.s. (mis)information news sources on facebook. In: Proceedings of the 21st ACM Internet Measurement Conference. IMC '21, pp. 444–463. Association for Computing Machinery, New York, NY, USA (2021). doi:10.1145/3487552.3487859. https://doi.org/10.1145/3487552.3487859
38. OECD: 21st-Century Readers: Developing Literacy Skills in a Digital World. OECD Publishing, `https://www.oecd-ilibrary.org/content/publication/a83d84cb-en` (2021)
39. Lee, J.: Size matters: Early vocabulary as a predictor of language and literacy competence. Applied Psycholinguistics **32**(1), 69 (2011)
40. Curtis, M.E.: The role of vocabulary instruction in adult basic education. Comings, J., Garner, B., Smith, C., Review of Adult Learning and Literacy **6**, 43–69 (2006)
41. Ouellette, G.P.: What's meaning got to do with it: The role of vocabulary in word reading and reading comprehension. Journal of educational psychology **98**(3), 554 (2006)
42. National Research Council: Improving Adult Literacy Instruction: Options for Practice and Research. National Academies Press, Washington, DC (2012). doi:10.17226/13242
43. Schmitt, N.: An Introduction to Applied Linguistics. Routledge, New York (2013)
44. Laurén, C., Nordman, M.: Special Language: From Humans Thinking to Thinking Machines. Multilingual Matters, Clevedon, Philadelphia (1989)
45. Nation, I.: How large a vocabulary is needed for reading and listening? Canadian modern language review **63**(1), 59–82 (2006)
46. Beglar, D., Nation, P.: A vocabulary size test. The language teacher **31**(7), 9–13 (2007)
47. Grave, É., Bojanowski, P., Gupta, P., Joulin, A., Mikolov, T.: Learning word vectors for 157 languages. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (2018)
48. Goldberg, Y., Orwant, J.: A dataset of syntactic-ngrams over time from a very large corpus of english books. In: Second Joint Conference on Lexical and Computational Semantics, Atlanta, Georgia, USA, pp. 241–247 (2013)
49. Satopaa, V., Albrecht, J., Irwin, D., Raghavan, B.: Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In: 2011 31st ICDCSW, pp. 166–171 (2011). IEEE
50. Antunes, M., Gomes, D., Aguiar, R.L.: Knee/elbow estimation based on first derivative threshold. In: 2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService), pp. 237–240 (2018). IEEE
51. Beasley, T.M., Erickson, S., Allison, D.B.: Rank-based inverse normal transformations are increasingly used, but are they merited? Behavior genetics **39**(5), 580 (2009)
52. Fatehkia, M., Kashyap, R., Weber, I.: Using Facebook ad data to track the global digital gender gap. World Development **107**, 189–209 (2018)
53. Roser, M., Ritchie, H., Ortiz-Ospina, E.: Internet. Our World in Data (2022). https://ourworldindata.org/internet
54. World Economic Forum: The global gender gap report. (2020). World Economic Forum Genebra
55. OECD: PISA 2018 Results (Volume II), p. 376 (2019). doi:10.1787/b5fd1b8f-en. https://www.oecd-ilibrary.org/content/publication/b5fd1b8f-en
56. United Nations Development Programme: Technical Notes: Calculating the human development indices—graphical presentation. Retrieved from UNDP website: `https://hdr.undp.org/sites/default/files/2021-22_HDR/hdr2021-22_technical_notes.pdf`. (Accessed on 10/26/2022) (2021)
57. Coppedge, M., Gerring, J., Knutsen, C.H., Krusell, J., Medzihorsky, J., Pernes, J., Skaaning, S.-E., Stepanova, N., Teorell, J., Tzelgov, E., *et al.*: The methodology of "varieties of democracy"(v-dem). Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique **143**(1), 107–133 (2019)
58. Warf, B.: Geographies of the Internet. Routledge, New York (2020)
59. World Bank: World development report 2016: Digital dividends. (2016). Washington, DC: World Bank
60. Zickuhr, K., Smith, A.: Digital differences. Retrieved December 7, 2020 from `https://www.pewresearch.org/internet/2012/04/13/digital-differences/` (2012)
61. Hilbert, M.: Digital gender divide or technologically empowered women in developing countries? a typical case of lies, damned lies, and statistics. Women's Studies International Forum **34**(6), 479–489 (2011). doi:10.1016/j.wsif.2011.07.001
62. Yamamoto, K., Khorramdel, L., Von Davier, M., et al.: Scaling PIAAC Cognitive Data
63. Facebook: Facebook Q2 2020 Earnings. `https://s21.q4cdn.com/399680738/files/doc_financials/2020/q2/Q2-2020-FB-Earnings-Presentation.pdf`.

(Accessed on 10/15/2020) (2020)

64. Magno, G., Weber, I.: International gender differences and gaps in online social networks. In: International Conference on Social Informatics, pp. 121–138 (2014). Springer
65. Kashyap, R., Verkroost, F.C.: Analysing global professional gender gaps using linkedin advertising data. EPJ Data Science **10**(1), 39 (2021)
66. Park, M., Thom, J., Mennicken, S., Cramer, H., Macy, M.: Global music streaming data reveal diurnal and seasonal patterns of affective preference. Nature human behaviour **3**(3), 230–236 (2019)
67. Masrai, A.: Vocabulary and reading comprehension revisited: Evidence for high-, mid-, and low-frequency vocabulary knowledge. Sage Open **9**(2), 2158244019845182 (2019)