Age and Gender in Reddit Commenting and Success

S. Craig Finlay *

Franklin D. Schurz Library
Indiana University South Bend, U.S.A.
E-mail: scfinlay@iusb.edu

ABSTRACT

Reddit is a large user generated content (USG) website in which users form common interest groups and submit links to external content or text posts of user-created content. The web site operates on a voting system whereby registered users can assign positive or negative ratings to both submitted content and comments made to submitted content. While Reddit is a pseudonymous site, with users creating usernames but providing no biographical data, an informal survey posted to a large shared interest community yielded 734 responses including age and gender of users. This provided a large amount of contextual biographical data with which to analyse user profiles at the first level of Computer Mediated Discourse Analysis (CMDA), articulated by Susan Herring. The results indicate that older Reddit users both formulate more complex writing and enjoy more success when rated by other users. Gender data was incomplete and as such only tentative results could be proposed in that regard.

Keywords: Computer mediated discourse analysis, CMDA, Reddit, Computer mediated communication, CMC, Age, Gender

1. INTRODUCTION

While a good deal of work has been done using Computer Mediated Discourse Analysis (CMDA) to determine differences in computer mediated communication behaviours as noted when looking at gender or structure within the communicative dynamic, the subject differences in age groups is largely unexplored. This is likely due to the difficulty of obtaining a large dataset wherein the users being

studied provide age information. When this has been done, such as in work by Kapidzic and Herring (2011) or Subrahmanyam and Greenfield (2004) in examining teen computer-mediated communication (CMC), the age context was made possible by studying teen chat rooms. While this approach provides researchers with a focused study on a particular age group, it also necessarily precludes likely participation by other age groups, preventing comparative analyses. Coupled with the fact that age is infrequently included in user profiles, the opportunity to look at how age differences manifest themselves in CMC is limited. This study makes use of a unique dataset: an informal request by a member of a message board within the website Reddit for age, gender, and nationality of message board subscribers. The resulting discussion thread included 734 responses. These responses were collected, and a sampling of comments from users was analysed according to the first level of CMDA (Herring, 2004). The first level of CMDA, corresponding to the four domains of language, is structure, with "phenomena [which] include the use of special typography or orthography, novel word formations and sentence structure" (Herring, 2004). The other three levels are listed as meaning, interaction, and social behaviour (Herring, 2004).

For this study, the following research questions were formulated:

R1: Is there an observable relationship between age and the phenomena classified by Herring (2004) as belonging to the structural level of CDMA, as quantifiable by examination of comment length, word length, and utterance length? If greater age may be associated with greater educational attainment and social development, one might expect to see an increase in complexity concurrent with advanced age.

R2: Is there a correlation between age and success of submitted links and comments as calculated by overall link and comment karma (the cumulative net total of link and comment scores, which are themselves the summation of a the Reddit voting system whereby other users rate the quality of submitted links and comments). Based on earlier analyses of Reddit comments, users seem to react more positively to longer, complex comments. Thus, there may also be a positive correlation between age and karma score.

Age as a topic of study is currently poorly under-stood in regard to effect on various aspects of CMD. While researchers have studied and learned about teen constructions of gender, sexuality, and overall usage, it is as yet unknown how and at what rate CMC evolves along with a user's age. The difficulty of studying CMC in context of gender and age is articulated by Androutsopolous (2006), who notes that user anonymity, coupled with the lack of the "main type of linguistic variable in the correlative paradigm … phonetics/phonology" results in fewer readily available dependent variables for researchers to analyse (p. 425). However, he notes that the diligent researcher can obtain this information, yielding textured, valuable results. This study is such an undertaking of computer-user biographical data which will hopefully shed light on issues of age and CMC. An important implication of this study is that there may be a reliable mechanism by which a user's age can be inferred purely from textual analysis. This should be of interest to CMC researchers as well as those engaged in advertising on the web. This study provides a foundation for important research into this area. If a system whereby age (or gender) can be inferred with a low error rate, the implications for CMC research could be enormous, as researchers would no longer have to manually mine biographical information.

2. REDDIT

Reddit is a popular user-generated content web site which allows users to submit links or original content to "subreddits," self-organized communities of interest created by users themselves. While membership within the site is not necessary to view posts, it is necessary to post links and comment on posted links. No content is hosted on Reddit itself. Links are usually either to news stories, or to images hosted on Imgur, a simple image sharing web site that generates short, random, persistent URLS for any uploaded image. Users may subscribe to these subreddits, and when signed in the web site recognizes user subscriptions and supplies popular content from those users' subscriptions. Some of these subreddits are quite general, such as "pics," a photo hosting board with (as of the time of writing), 1.28 million subscribers. Some subreddits more unique to Reddit include "todayilearned," with approx-

imately 800,000 subscribers. Users post interesting and obscure facts to this subreddit, leading to the common acronym "TIL," used throughout the web site. Another is the "F7U12," or "rage comic" subreddit, in which readers create simple comic strips using Reddit's own drag-and-drop template to tell personal amusing stories. The data for this survey was taken from r/atheism, a message board of atheists from around the world that describes itself as the largest such community on the internet.

A unique aspect of Reddit is the option to provide a simple assessment of approval or disapproval on any submitted content or on any comment posted in response to content (or to an existing comment). Known as "upvoting" or "downvoting," the system is democratic in that a user may only vote once for any given link or comment. Thus, a user cannot repeatedly up or downvote a post, though a user may replace an upvote with a downvote if so desired. The net ratio of upvotes minus downvotes is displayed next to a link or comment as a "score." That score is automatically assigned to a user's profile as that user's "karma." Users have two different karma ratings for submitted links and submitted content. Within the community, high karma scores are seen as status symbols, and may be seen, from the researcher's perspective, to be indicative of a Redditor's success on the site. Since most comments are not highly upvoted, comment karma, a cumulative score of many comments with one or two points apiece, may also be said to correlate to a user's activity level. Because of the importance Redditors place on karma scores, they have important implications for CMDA when used to study communication on this site.

3. LITERATURE REVIEW

Computer mediated discourse analysis (CMDA) is broadly outlined in Herring (2004), which explained the approach to studying computer mediated communication (CMC) as "applying to four domains or levels of language, ranging prototypically from smallest to largest linguistic unit of analysis: 1) structure, 2) meaning, 3) interaction, and 4) social behaviour" (p. 3). The first level, structure, focuses on word usage, sentence structure, typography, and orthography. It is this level

which concerns the present study. As the first step toward studying this data set, establishing the structural characteristics is a prerequisite toward more qualitative analyses.

A number of studies have looked at specific age groups and their use of computer mediated communication. Suzuki and Calzo (2004) looked at teens who sought advice on sex and health from two message boards. The study collected 273 questions and responses and found that teens were willing to ask questions regarding sex and sexual health in an online setting that they were unwilling to ask face-to-face from an adult. Gross (2004) looked broadly at online behaviour, using surveys. Gross found that boys and girls alike tended to use CMC to communicate on intimate topics with friends whom they knew outside of their online lives. Subrahmanyam, Greenfield, and Tynes (2004) looked at teen CMC and found that adolescents were taking advantage of the "screen" provided by the medium as a way to engage with others on topics of sexuality and gender, as part of their developmental processes. They also found that adolescents utilized CMC as a way to "practice" participating in different kinds of relationships with others.

The wealth of information provided by mining comment boards and collaborative websites has recently allowed researchers to look at various aspects of human communication. Ioannou (2011) examined how wikis facilitate collaborative creation by web users. Three researchers recently looked at comment ratings on message boards to examine the spread of ideas. Koteyko, Jaspal and Nerlich (2013) looked at user comments on UK tabloid news sites to examine evolving attitudes toward the climate change debate. Chiluwa (2009) innovatively applied CMDA to study the popularly termed "419" emails – hoaxes which attempt to convince recipients to pay money by informing them of large lottery winnings or other large sums of money that may be claimed.

A more recent look at gender was undertaken by Kapidzic and Herring (2011), who looked at gender and CMC in English-language teen chat sites. They found that gender differences manifest on a number of levels. Males were found to participate in more "invite" acts, requesting behaviour of females, whereas female users tended to react to those assertive behaviours. They also found that male communications were often

more overtly sexual in nature than those of their female counterparts. The study also looked at differences in self-representation in terms of posted profile photos. Herring (2010) looked at the concept of the "floor" in regard to gender, specifically in terms of success in gaining responses and controlling the floor of the conversation. That study found that males were more successful in garnering responses in message boards, probably due to higher rates of message posting. Such is of particular importance to the present study in that the dataset provides an opportunity to determine "success" as a Redditor (as evidenced by karma score), in the context of gender. In addition, such an approach may equally be applied to age.

The wealth of data (tens of thousands of comments) makes possible other analyses of writing not attempted here. For example, Herring and Paolillo (2006) looked at the "gender" of weblog genres using the Gender Genie program. In regard to the program itself, the researchers found it had mixed success in identifying author gender. The concept of automatically identifying age is an ultimate goal of this project, and the massive potential dataset offered by the Reddit interface providing all comments ever posted by each user means substantial steps could be taken toward this goal. In short, the large amount of work done identifying the ways in which gender is manifested in CMC provide a blueprint for studying age. Much of the work done in regard to the former focus on notions of power (see Herring, 2003) and gender, specifically manifestations of sexism on the internet. Obviously this dynamic would not manifest itself as such in the interactions of different age groups. However, the social position occupied by certain age groups, such as middle-school students as opposed to working adults, may well influence interactions in an online environment.

4. METHODOLOGY AND DATA COLLECTION

The methodology of analysis of this data set is CMDA, specifically the first level of CMDA as outlined in Herring (2004). The dataset for this study was collected from an informal survey posted to the r/atheism subreddit on December 17, 2011. User "xtimrs" posted to r/atheism a message wondering as to the demographic makeup of the subreddit, which at the time had a little less than 500,000 subscribed readers. While the original post simply asked for suggestions on workshopping a possible demographic survey, users began responding with largely "age, sex, location" comments. Of the 998 comments in the threads, 734 responded to the survey. This information was collected manually. While age is the primary focus of this study, there was a sufficient population of female respondents to enable a gender-based analysis of submitted comments [Table 1].

Gender, along with age and location data, was manually collected and assembled into a spreadsheet and then used as the basis for further data collection, starting with manual collection of user comments and link karma. Recall that comment karma may be broadly associated with user activity, as most comments are not highly upvoted and therefore a user's comment karma score grows slowly, over time. Also, as high karma scores are seen as status symbols, they may also be said to indicate a user's "success" as a Redditor. Comment and link kar-

Table 1. Self-Reported Gender of Reddit Users in r/atheism

Gender	Number	Percent
Male	470	64.03%
Female	122	16.62%
No response	141	19.21%
Other	1	0.14%
Total	734	

ma for each responding user was collected manually by clicking on each responding user's name in turn.

Among the publicly available information on Reddit is past activity by a user. Clicking on a user's name will display a list of every comment and link that user has ever posted to Reddit. Initially, it was envisioned that an automated crawler would be constructed that could gather this data in its entirety. Such a crawler could not be constructed in time, and as such comment collection was also manual. This meant that complete collection was impossible, as for most users the number of comments posted is likely to be in the dozens and often in the hundreds. To exhaustively document all activity by the 734 survey respondents would mean collecting tens of thousands of comments and links and their associated metadata (posting time, scores, etc.). It was thus decided to collect a sample of comments from users across all age groups, thereby enabling comparative analysis, if not presenting a necessarily accurate picture of the community as a whole. To do this, the first page of comments by each responding user was manually copied and pasted into an Excel sheet. Clicking a user's name brings up all comments by that user, divided into pages. To do this, bins were first created which may be said to roughly correspond to certain commonly accepted educational developmental periods [Table 2].

The logic behind this that Reddit is a U.S. based website and the ages in these bins commonly correspond in the United States to certain stages of educational prog-

ress. Those individuals in group 1 would be expected to be in grades 6-10 (middle school through mid-high school), group 2 to high school seniors, group 3 to college undergrads, group 4 to recent college grads, those entering the workforce and those in graduate school, 5 to young adults settling into professional career paths and 6 to adults firmly established in the workforce. Group 6 is by far the largest in terms of timespan, covering ages from 36 to 67. It is also the smallest group. For each group, the average comment and link karma was calculated, to provide a baseline for the "average" user of that demographic [Table 3]. Then, for each age group, the first five users with comment karma above and below this baseline were selected as that demographic's representative age group. This population of 60 was later expanded to 90 to include more female users, allowing for another level of comparative analysis. In all, data for 30 female users was collected. For each of these users, the most recent five comments posted were collected for a dataset of 450 comments. After collection, comments were analysed by manually counting all comments for number of utterances. Additionally, content was analysed with the Microsoft Word spelling and grammar check to determine average word length and readability statistics. Further studies will be conducted utilizing a crawler to automate data collection, enabling broader conclusions. The present results should be considered a compass pointing out future research directions.

Table 2. Age Bins of Reddit Survey Respondents

Group	Age	Number	Percent
1	12-16	120	16.35%
2	17-18	114	15.53%
3	19-22	204	27.79%
4	23-27	150	20.44%
5	28-35	83	11.31%
6	36+	59	8.04%
	No response	4	0.54%
	Total	734	100%

Table 3. Average Link and Comment Karma, By Age Bin

Age	Number	Average link karma	Average comment karma
12-16	120	321	649
17-18	114	534	1280
19-22	204	586	1307
23-27	150	497	2342
28-35	83	356	2979
36+	59	569	3006
No response	4	4	N/A
Overall	734	508	1699

5. RESULTS

The average Reddit r/atheism user may be described as a male in his late teens to mid-20s. Female users are very much in the minority, at some 16.6% of respondents. However, 19% of respondents neglected to include age in their comments, and as such the gender demographics of the Reddit r/atheism population may vary some from the proportions seen here. Far more complete was the age data, with only 4 users neglecting to provide information. We can see in Table 2 that a majority of Redditors fall within the 12-16 (16.3%), 17-18 (15.5%), 19-22 (27.7%) and 23-27 (20.4%) age groups. Users over the age of 28 are a distinct minority, although enough data was collected to enable meaningful analysis. In Table 3 it is apparent that there is a positive correlation between user age and comment karma score. Each age group revealed higher average comment karma than the group below it, although the difference between the 28-35 and 36+ groupings is minimal. The relationship is especially striking when visualized in bar graph form [Figure 1].
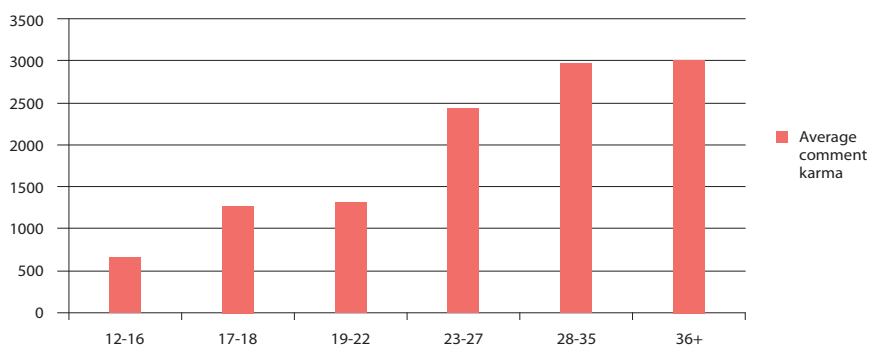
Age differences in Reddit karma score

Fig. 1 Age differences in Reddit karma score

The average karma of the two highest age groups is so similar as to be functionally identical for the purposes of this study (2979 and 3006). Together, they have a population of 142, making them more or less equal in size as a demographic (albeit a very, very diverse one), to the other age bins. It is striking that the average karma score for this highest age bracket (or brackets) is on average over 4.5 times higher than for those in the youngest age bracket. The 12-16 year old age bracket displayed an average comment karma of only 649, a little over half that of the next age bracket, the 17-18 year olds, which displayed an average karma of 1280. This total was very similar to the 19-22 year old group, with an average karma score of 1307. Interestingly, this is identical in size to the gap between the two highest age groups. There is a large gap between the 19-22 year olds and the next age group, the 23-27 year olds, which had an average karma score of 2342. No substantial differences were observed in link karma among the age groups.

Comment karma was also examined in the context of gender [Table 4]. While male respondents were found to have a higher comment karma score (1737) than female respondents (1257), both were lower than the average score computed for the 141 who declined to include gender information. It must be assumed that the inclusion of missing gender data would alter the results and as such these numbers should be taken with the proverbial grain of salt.

When the average length of comments was analysed, age and gender differences were apparent as well [Table 5]. The average overall comment length was 28.1 words. While not as striking as the differences in user karma, a similar pattern is evident, with higher age groups writing longer comments. The 12-16 and 17-18 age groups wrote similar-length posts, 22.8 and 23, respectively. Users aged 19-22 years old wrote com-

Table 4. Karma in the Context of Self-Reported Gender

Gender	Number	Average link karma	Average comment karma
Male	470	517	1737
Female	122	352	1257
No response	141	493	1944
Other	1	N/A	N/A
Total	734	508	1699

Table 5. Average Comment Length by User Age (Number of Comments per Age bin = 75)

Age	Average comment length
12-16	22.8
17-18	23
19-22	26
23-27	32.1
28-35	32.9
36+	32.2
Overall	28.1

ments 26 words long on average. The 23-27, 28-35 and 36+ age groups all wrote comments, on average, of 32 words long, with decimal-size variation. When averaged together, the two highest age groups write around 10 more words per comment than the two smallest age groups.

When examined according to age, the length of the average comment by males was found to be similar to the two smallest age groups, at 23.3 [Table 6]. Females, by contrast, wrote posts similar in length to the three highest age groups, at 32.4. Again the group which neglected to provide gender information had the highest result, with an average number of 34.5, greater than any single age group.

To consider a unit of writing utterance in this study was it is a clause, as articulated by Condon and Cech (1996). The number of utterances per comment for the sampled data was manually counted and revealed to be 4.7, with an average overall utterance length of 5.8 words [Table 7]. While the average utterance length did not vary widely among age groups, the number of utterances per comment increased according to age. This follows when considering the longer average posts by those in the higher age brackets. Average word length was found to be broadly similar across all age groups.

The same analysis was applied to gender, and it was found that male Redditors write fewer utterances (4.7) and shorter utterances (4.9 words) than their female counterparts (5.4 and 6 words). Those users who neglected to provide gender data displayed a slightly higher number of average utterances (5.6), and identical words per utterance as the female respondents [Table 8].

Table 6. Average Comment Length by Reported Gender

Gender	Number	Average comment length
Male	470	23.3
Female	122	32.4
No response	141	34.5

Table 7. Utterances per Comment, Words per Utterance, and Average Word Length, By Age Group

Age	Number	Utterance/comment	Words/utterance	Word length
12-16	120	3.9	5.8	4.3
17-18	114	3.7	6.1	4.3
19-22	204	4.2	6.1	4.2
23-27	150	5.4	5.9	4.3
28-35	83	5.8	5.65	4.3
36+	59	5.6	5.6	4.2
No response	4	N/A	N/A	N/A
	734	4.7	5.8	

25

Table 8. Utterances Examined by Gender Identification

Gender	Number	Utterance/comment	Words/utterance	Word length
Male	470	4.7	4.9	4.3
Female	122	5.4	6	4.2
No response	141	5.6	6	4.2

6. DISCUSSION

In an earlier study on Reddit comments, looking at 94 comments posted to a single discussion thread in response to a rage comic in the F7U12 subreddit, users were arbitrarily binned according to comment karma, with length of comments and utterance data analysed in that context [Table 9]. That study found that users in the higher karma brackets tended to write longer posts with a greater number of utterances, and with longer utterances, than users with lower average karma. While the numbers analysed in that study were too small to draw more than preliminary conclusions, those would appear to have been confirmed by this study. Users with higher karma scores tend to write longer comments. Whether this is because longer comments are better received, or because more experienced users feel more comfortable writing longer comments cannot be determined at this time, and no causality should be assumed.

To recap the research questions formulated above:

R1: Is there an observable relationship between age and the phenomena classified by Herring (2004) as belonging to the structural level of CDMA, as quantifiable by examination of comment length, word length, and utterance length? If greater age may be associated with greater educational attainment and social development, one might expect to see an increase in complexity concurrent with advanced age.

R2: Is there a correlation between age and success of submitted links and comments as calculated by overall link and comment karma (the cumulative net total of link and comment scores, which are themselves the summation of a the Reddit voting system whereby other users rate the quality of submitted links and comments). Based on earlier analyses of Reddit comments, users seem to react more positively to longer, complex comments. Thus, there may also be a positive correlation between age and karma score.

An important result of this study is the identification of differences in computer mediated communication among individuals of different age groups. Consistently, younger users have lower karma scores and write shorter posts than older users. They also write less complex comments, with a smaller number of utterances. This could indicate that as individuals attain more education their confidence and writing ability increases and this is manifested in their online communications. Therefore, R1 and R2 are answered in the affirmative: a positive correlation exists be-

Table 9. A Previous Study Examining Commenting Behaviour According to Karma Score

Population (by user comment karma)	Words per comment	characters per word	number of utterances	words/utterance
bottom third (n=25)	14.3	4.1	2.07	6.9
middle third (n=24)	15.41	4	2.56	6.01
top third (n=24)	21.82	4.1	2.68	8.147

tween success of submitted comments and age, and age differences manifest themselves at least at the level of comment length. It is interesting that no striking differences exist between age groups in terms of link karma, the aggregate score of all links submitted to Reddit. This would indicate that age has no bearing on the success of submitted content. This may be because the users' increased intellectual ability has less opportunity for expression in the short description space allowed by Reddit when submitting a link than in the comparatively freer platform of commenting.

7. LIMITATIONS

It is disappointing that the gender data was so incomplete, and the type of analysis conducted in Herring (2010) could not be repeated here. That more respondents neglected to include gender data than self-identified as female likely renders any conclusions tentative at best. Evidence of this may be seen in the fact that males had higher average karma scores but much shorter average comments. The lower number of utterances and shorter utterances per comment for males also indicates a decreased complexity of comments. However, those individuals who did not respond with gender information had both higher comment karma and longer average comments than either males or females, which seems to confirm earlier conclusions. More data needs to be collected to enable drawing a conclusion one way or the other.

8. FUTURE WORK

Much more work remains to be done with this dataset. The first level of CMDA was not exhausted here, not to mention issues of topicality, interactivity, or politeness. It has been established that at least some age-based differences are manifested in online communication. With further research, other differences will undoubtedly surface. The next task must be to exhaustively collect all comments and links submitted by all users within the dataset. Then researchers may begin the long task of applying higher levels of CMDA to what will be a very large corpus of data.

REFERENCES

Chiluwa, I. (2009). The discourse of digital deceptions and '419' emails. *Discourse Studies, 11*(6), 635-660. doi:10.1177/1461445609347229

Condon, S., & Cech, C. (1996). Discourse management strategies in face-to-face and computer mediated decision making interactions. *The Electronic Journal of Communication, 6*(3). Retrieved from http://www.cios.org/EJCPUBLIC/006/3/006314.HTML.

Gross, E. (2004). Adolescent internet use: What we expect, what teens report. *Applied Developmental Psychology, 25*, 633-649.

Herring, S. C. (2010). Who's got the floor in computer-mediated conversation? Edelsky's gender patterns revisited. *Language@Internet*, 7, article 8. Retrieved from http://www.languageatinternet.org/articles/2010/2857.

Herring, S. C., & Paolillo, J. C. (2006). Gender and genre variation in weblogs. *Journal of Sociolinguistics, 10*(4), 439-459. Preprint retrieved from http://ella.slis.indiana.edu/~herring/jslx.pdf.

Herring, S. C., & Martinson, A. (2004). Assessing gender authenticity in computer-mediated language use: Evidence from an identity game. *Journal of Language and Social Psychology, 23*(4), 424-446. doi: 10.1177/0261927X04269586

Herring, S. C. (2004). Computer-mediated discourse analysis: An approach to researching online behaviour. In S. A. Barab, R. Kling, & J. H. Gray (Eds.), *Designing for Virtual Communities in the Service of Learning* (pp. 338-376). New York: Cambridge University Press. Preprint retrieved from http://ella.slis.indiana.edu/~herring/cmda.pdf.

Herring, S. C. (2003). Gender and power in online communication. In J. Holmes & M. Meyerhoff (Eds.), *The Handbook of Language and Gender* (pp. 202-228). Oxford: Blackwell Publishers. Retrieved from http://ella.slis.indiana.edu/~herring/gender.power.pdf.

Ioannou, A. (2011). Online collaborative learning: The promise of wikis. *International Journal Of Instructional Media, 38*(3), 213-223.

Kapidzic, S., & Herring, S. C. (2011). Gender, communication, and self-presentation in teen chatrooms revisited: Have patterns changed? *Journal of Computer-Mediated Communication, 17*(1), 39-

59. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.2011.01561.x/full.

Koteyko, N., Jaspal, R., & Nerlich, B. (2013). Climate change and 'Climategate' in online reader comments: A mixed methods study. *Geographical Journal, 179*(1), 74-86.

Subrahmanyam, K., Greenfield, P., & Tynes, B. (2004). Constructing sexuality and identity in an online teen chat room. *Applied Developmental Psychology, 25*, 651-666.

Suzuki, L., & Calzo, J. (2004). The search for peer advice in cyberspace: An examination. *Applied Developmental Psychology, 25*, 685-698.