Contents lists available at ScienceDirect

Heliyon

journal homepage: www.heliyon.com

We tweet Arabic; I tweet English: self-concept, language and social media

Justin Thomas [a,*], Aamna Al-Shehhi [b], Marwa Al-Ameri [a], Ian Grey [c]

[a] *Zayed University, United Arab Emirates*
[b] *Khalifa University, United Arab Emirates*
[c] *Lebanese American University, Lebanon*

ARTICLE INFO

Keywords:
Psychology
Twitter
Arabic
Self concept
Social media
Culture
UAE

ABSTRACT

Differences in self-concept have been observed across cultures. Participants from collectivist societies tend to describe themselves using social and relational attributes (mother, student, Arab) more frequently than their individualist counterparts, who tend to rely more heavily on personal attributes (fun, tall, beautiful). Much of this past research has relied on relatively small samples of college students, tasked with spontaneously reporting self-concepts in classroom settings. The present study re-examines these ideas using data extracted from Twitter, the popular social media platform. In analysis one, the Twitter biographies of individuals exclusively posting messages in English ($N = 500$) and those posting only in Arabic ($N = 500$) were content analyzed and quantified for differences in the frequency of personal versus social attribute use. Analysis two applied a bilingual word counting algorithm to the biographies of a larger sample of Twitter users ($N = 242{,}162$), exploring the relative frequency of social attributes, specifically familial roles (e.g. mother, father, daughter, son), across both English and Arabic users. In analysis one, the Twitter biographies of exclusive Arabic users contained significantly more social attributes than their English using counterparts. In analysis two, Arabic biographies contained significantly more familial references than their English language counterparts. These findings support the idea that cultural values may influence self-construal. Big data extracted from social media platforms appear to offer a useful means of exploring self-concept across cultures and languages.

1. Introduction

Self-concept represents the sum of an individual's beliefs about their attributes and who and what the self is (Baumeister, 1998). Cross-cultural studies of self-concept typically contrast allocentric or collectivistic self-conceptions with more individualistic and idiocentric ones (Bond and Cheung, 1983; Heine, 2001; Ma and Schoeneman, 1997). Cultural values are thought to influence self-construal, giving rise to differences in the relative emphasis placed on personal attributes (e.g. personal traits and achievements) and social attributes (e.g. relationships and group memberships) when self-concept is elicited (Heine, 2001). Within relatively individualistic societies (e.g. USA, UK, Australia), the value placed on independence and standing out is held to lead to the development of self-concepts that give relatively greater priority to personal attributes. Conversely, within collectivist societies (e.g. Japan, Korea, India), the higher value assigned to interdependence and fitting in is thought to lead to self-concepts that place relatively greater emphasis on social attributes (Markus and Kitayama, 1991).

This idea that cultural values influence self-construal is qualified by the observation that personal attributes tend to feature more prominently than social attributes across all cultural groups. In other words, even people socialised within collectivist societies tend to mention relatively more personal attributes than social ones in tasks designed to elicit self-concept, albeit at a lower ratio than their individualistic counterparts. This idea is known as individual-self primacy and suggests that personal attributes will typically be more salient than social attributes regardless of cultural orientation (Gaertner et al., 1999).

Studies exploring self-concept across cultures generally support the idea that cultural values influence self-concept, and that personal attributes tend to be more frequently mentioned than social ones. In a study spanning seven countries, Watkins et al. (2003), as hypothesised, found that differences in the relative percentage of self-descriptors classified as individualistic/idiocentric (e.g. autonomous traits and personal preferences) or collectivistic/allocentric (e.g. group memberships and social relationships) varied by collectivism and individualism. Similarly, comparing students from India and the USA, Dhawan et al. (1995) found that American students were significantly more likely to stress individual rather than group identity, while Indian students demonstrated the

[*] Corresponding author.
E-mail address: Justin.thomas@zu.ac.ae (J. Thomas).

https://doi.org/10.1016/j.heliyon.2019.e02087
Received 13 November 2018; Received in revised form 3 June 2019; Accepted 10 July 2019

opposite pattern. In a study comparing US and Kenyan college students, Ma and Schoeneman (1997) found the hypothesised preponderance of allocentric (social/collectivist) responses among the Kenyan students relative to their US counterparts. This study also explored western acculturation, finding that the relative frequency of allocentric (social) descriptors was greatest among the more traditional - less western acculturated - Kenyans.

This distinction between relatively allocentric self-concepts and idiocentric ones is particularly well supported by studies comparing East Asian (e.g. Japanese, Korean, Taiwanese) participants with their North American counterparts. Such studies find that participants from the collectivist societies tend to mention personal attributes less frequently (Bond and Cheung, 1983; Cousins, 1989; Kanagawa et al., 2001; Rhee et al., 1995). Not all studies, however, fully support the cultural influence hypothesis. Several studies report only partial support (Kanagawa et al., 2001), and others offer no support at all (Watkins et al., 1998). In a review of such studies, del Prado et al. (2007) suggest that virtually all of the cultural influence studies explore only the individualism/collectivism dimension of national culture and that there might be other dimensions of culture that influence or moderate the outcomes. Additionally, relatively small sample sizes and reliance upon a single method of data collection and analysis - the Twenty Statements Task (Kuhn and McPartland, 1954) - might also account for the minor equivocation. We could also add to this list of limitations, that the majority of previous research in this area has relied upon student samples, which, as in the case of the Kenyan study (Ma and Schoeneman, 1997), can represent a more highly educated and western acculturated demographic.

The present study introduces a novel method of exploring self-concept, which simultaneously addresses the previous issues of small samples sizes and the over-reliance on student samples. Specifically, the present study examines the idea of cultural influences on self-concept using a large dataset from Twitter, a popular social media platform.

Large data sets (big data) associated with search engines and social media platforms such as Twitter and Facebook are increasingly viewed as a means of exploring psychological variables (Yang and Srinivasan, 2016). For example, Dodds et al. (2011) used Twitter data to quantify the temporal patterns of happiness. Using an algorithm based on frequency counts of positive and negative words, they identified the days, weeks and hours of the day that were most strongly associated with written expressions of positive and negative affect. Similarly, Yang and Srinivasan (2016) used an algorithmic analysis of a large Twitter dataset to explore life satisfaction based on the items of a well-validated subjective well-being measure. Other studies have used big data to explore a variety of other domains, from cultural values (Garcia-Gavilanes et al., 2013) to help-seeking behaviours related to influenza epidemics (Ginsberg et al., 2009). These big data analyses typically find patterns in the data that are consistent with intuition, and concordant with previous findings based on more traditional research methods. Additionally, unlike self-report surveys, social media datasets can claim a degree of ecological validity, being free from research related demand characteristics at the point of data collection. There is an emerging consensus that large social media datasets can be a useful adjunct to traditional research methods (Yang and Srinivasan, 2016).

The use of such datasets, however, comes with some inherent limitations. It can be difficult to ascertain with certainty demographic attributes such as gender, age and nationality etc. In the present study, for example, cultural orientation is inferred from language use, and gender from display names. Big data analysis, in general, relies heavily on such heuristics - rules of thumb. There is an acceptance, however, that this occasional lack of precision is compensated for by the volume of the data involved (Dodds et al., 2011).

The present study attempts explore self-concept from the biographies associated with Twitter accounts. Commonly known as a Twitter bio, this summary field (160 characters maximum) is frequently used to introduce the account owner to other Twitter users. The author of the Twitter bio will often use the space to say something about themselves. Many Twitter bios undoubtedly represent brief expressions of self-concept, freely volunteered outside of a laboratory or experimental setting. In addition to enhancing ecological validity, the use of Twitter data also facilitates a far larger and more occupationally diverse sample than is typically found in paper-based explorations of self-concept. One challenge of the current methodology, however, is being able to assign demographic and grouping variables; age, gender and cultural orientation must all be inferred. In the present study, exclusive language use (Arabic versus English) is taken as an imperfect proxy for cultural orientation (collectivism versus individualism), while we derive gender from the user's display name.

This study involves two main analyses, the first explores the Twitter bios of 1000 Twitter users, comparing the relative frequency of personal versus social attributes between Twitter users who tweet exclusively in English and those tweeting exclusively in Arabic. It is predicted that the bios of exclusive Arabic users will be associated with more frequent mentions of social attributes, and the bios of those tweeting exclusively in English will be associated with relatively more mentions of personal attributes. It is also predicted that personal attributes will be more frequently mentioned than social attributes across both language groups (individual self-primacy). The second analysis will use a larger dataset - 242,162 Twitter biographies - and bilingual word count algorithm. The algorithm will search all English and Arabic twitter bios for familial references (e.g. Mother father, son, daughter). Previous self-concept research refereing to family-role is typically classed as a social/relational attribute (del Prado et al., 2007). It is predicted that Arabic twitter bios will make more frequent familial references.

2. Method

2.1. Dataset description

The UAE Twitter dataset for 2015 was a random sample of the entire corpus, numbering 8.2 million tweets collected between (January 1, 2015, to January 1, 2016). The data obtained can be classified into two main categories; features related to the user and features related to the text (tweet). User features included display name, account summary (Twitter bio) and language. Text level features included, text language, geolocation, location name, and posted time. There are 24 different languages; 44% of tweets were Arabic and 39% English. These were the two most commonly used languages in the data set. The diversity in the languages reflects the UAE's diverse expatriate population, many of whom do not speak Arabic (Thomas, 2014). Table 1 summarizes additional information concerning the data set.

The Twitter accounts, exclusively tweeting in either English or Arabic, were computationally randomised. The first 1000 eligible accounts (500 Arabic users and 500 English users) were extracted for analysis. The fields included in this subsample were: display name language and the account summary referred to hereafter as the Twitter bio. The Twitter bio is an optional field and can be left blank; it is also limited to a maximum of 160 characters. Exclusion criteria included accounts that related to organizations (businesses, newspapers, information services etc.), and accounts with blank Twitter bios. Table 2 Details a small sample of Twitter bios from the study.

Table 1
Breakdown of language use and unique users from the UAE Twitter data set for 2015.

Language	Number of tweets	Unique users
Arabic	3,126,163	58,776
English	2,816,777	124,543
Other	2,262,602	6,175
Total	8,205,542	189,494

Table 2
Examples of Twitter bios from the UAE Twitter data set for 2015.

Example Twitter bios
Sports entrepreneur, influencer citizen of the world, feminist, baseball, dogs and art deco
I'm a simple person who hides a thousand feelings behind the happiest smile
A son, brother, cousin, poet, biker, rafter, and a coffee addict, ...
Zayed University student.
27-year-old Scorpio studied criminal sociology. Intrigued by people's thoughts and societal behaviour. A member of breathing numbers

2.2. *The coding categories for the twitter bios*

The coding categories (personal versus social attributes) were based on the work of del Prado et al. (2007), which comprises the following 11 categories: traits, social identities, preferences, aspirations, activities, attitudes, skills, physical descriptions, emotional states, individuating self-references and unclassifiable. Each of these categories was further subdivided, and prototypical examples were provided to assist the coders. For example, the social identity category included the following subcategories and prototypes:

a. Social role-status (student, major)
b. Family role-status (I am a daughter. I am the last child.)
c. Family information (I have a brother. I am close to my family.)
d. Social relationships (I am a friend of.)
e. Social information (I have a wife. I am in a club. I have many friends)
f. Ethnicity/race/nationality (Emirati, Londoner, Khaleeji)
g. Gender (boy, woman)
h. Self-ascribed identities (poet, hunter, feminist.)
i. Origin (from the UK)
j. Religion (Muslim, Quranic Verse)
k. Occupation (Salesperson, Engineer)
l. Denial of social identity (not a typical Christian, not close to my family)
m. Universal-oceanic (human being, earthling)

(Adapted from del Prado et al., 2007)

2.3. *Coding procedure*

The coding categories and prototypes were laid out in a Microsoft Excel spreadsheet, with the first column containing the Twitter bio and the subsequent columns reflecting the coding matrix with accompanying prototypes/examples. Two raters independently tallied instances of each of the 11 categories, awarding a score of 1 for each instance identified in a Twitter bio. After the first pass, inter-rater reliability was acceptable (r [998] $=$ .81). The few discrepancies in coding were resolved through review and discussion, with the first author of the paper providing adjudication. Ultimately, social categories and personal categories were summed to provide each Twitter user with a total social attribute score and a total personal attribute score. Not all Twitter biographies could be categorised. Uncategorized biographies were scored as zero.

2.4. *Analysis two*

A second analysis was undertaken using the above-mentioned dataset. Analysis two was automated and applied to all English and Arabic bios, using a bilingual word counting algorithm. The target words were first-degree familial references (listed in table two). Past research exploring self-concept has classed mentions of familial role (e.g. mother of, son of etc.) as social attributes (del Prado et al., 2007). This category attribute is also directly translatable across both languages.

This study, analysis one and two, was exempted from applying for ethical clearance by the research ethics committee at Zayed University. This exemption was based on the study's use of anonymized and publicly available data sources.

English	Arabic
Mom, mother, mum	أم، ماما
Dad, father	أب، بابا
Brother, sister	أخ، أخت
Son, daughter	أبن، بنت

3. Results

Twitter bios are limited to 160 characters; in the present dataset the mean number of words per bio was 8.74 ($SD = 7.04$). The most frequently identified personal category was individuating self-references, at 418 instances. High-frequency social subcategories were: religion (263 references), nationality/ethnicity (120) and occupation (111). Overall, 852 personal attributes were detected, compared with 712 social attributes. The mean number of references to personal attributes in Twitter bios was .67 ($SD = .66$), which was greater than the mean for social attributes .59 ($SD = .70$). Using a paired samples t-test, this difference was statistically significance (t [946] $= 1.91$, $p = 0.47$, $d = 0.11$). Comparing personal and social attribute frequency between languages, the mean score for personal attributes among English users was .85 (SD $=$ .69) while for Arabic users it was .47 (SD $=$ .58). This difference was statistically significant (t[945] $= 8.66$, $p < .001$, $d = 0.59$). Conversely, social attribute frequency was greater among Arabic users ($M = .79$, $SD = .67$) compared with English users ($M = .41$, $SD = .64$). This difference was also statistically significant (t[945] $=$ -8.50, $p < .001$, $d = 0.58$.). Fig. 1, below, depicts this contrasting pattern of results.

Further analysis involved categorising users as either social or personal-attribute dominant. We achieved this by subtracting each user's social attribute score from their personal attribute score (the total number of social attributes mentioned - the total number of personal attributes mentioned). After the subtraction, those with positive scores were categorised as personal-attribute dominant, while those with negative scores were classified as social-attribute dominant. Those with equal scores for both categories, or no score at all, were excluded from the analysis. This categorical analysis revealed that among Arabic language users ($N = 408$), 58.3 % were social-attribute dominant. Whereas among the English language users ($N = 439$), 64.8% were personal-attribute dominant. Using Pearson's Chi-Square test, these differences were statistically significant $\chi^2(1, N = 847) = 44.59, p < .001$.

In light of previous self-concept research reporting females as being more likely to use social attributes (Jackson et al., 1994), we also analysed the Twitter bios by gender. To do this, we identified users with display names that lent themselves to unequivocal gender categorisation. We only included accounts with gender-specific forenames (Muhammad99, Mariam_123, SweetJuliet86 etc.) in the analysis. The gender classification process resulted in identifying 334 males, and 342 females. In terms of personal attributes, females (M $=$ .71, SD $=$ .71) used slightly more than males (M $=$ .65, SD $=$ .64), and the same pattern was also true for social attributes, M $=$ .61, SD $=$ .66 and M $=$ .56, SD $=$ .68, respectively. Categorical analysis (as detailed above) revealed that, compared with males, females were more likely than males to be categorised as social-attribute dominant (55.1%) and less likely to be categorised as personal-attribute dominant (47.5 %). This categorical gender difference was statistically significant, $\chi^2(1, N = 676) = 3.76, p = .027$.

3.1. *Analysis 2*

Looking at the use of familial references across all English and Arabic Twitter bios revealed that Arabic users were significantly more likely to

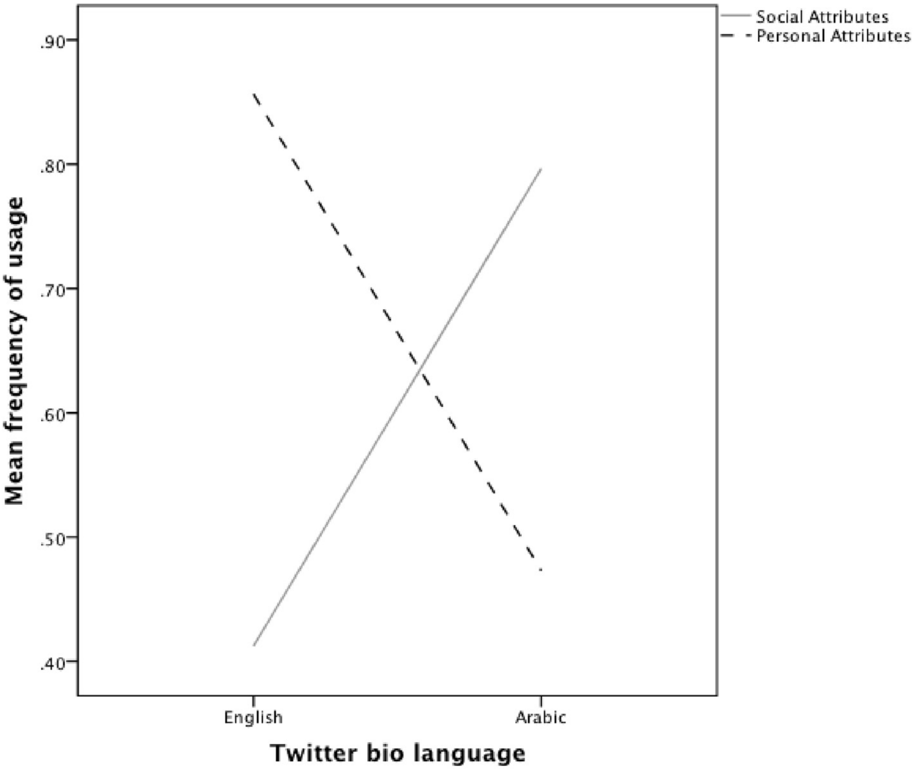

Fig. 1. Graph plotting the means for the frequency of social and personal attribute mentions in the Twitter bios of English and Arabic users.

mention family roles in their biographies $\chi^2(1, N = 236{,}819) = 505.58, p < .001$. With examples including “Mother, Grandmother”.

	Familial reference	No Familial reference
Arabic	3.34% (2357)	96.66% (70663)
English	1.8% (2986)	98.20% (166156)

4. Discussion

The present findings support both the idea that culture may influence self-construal, and that personal attributes are more frequently mentioned than social ones (individual-self primacy). Arabic users mentioned relatively more social attributes in their Twitter bios than did their English using counterparts. Conversely, English users provided relatively more personal attributes compared to Arabic users. Both of these language-related results were highly significant with medium effect sizes, suggesting fairly robust differences between the two groups of language users. The use of language as a proxy for culture offers only heuristic value at best and this lack of verifiable demographic information is an inherent constraint associated with the use of most social media datasets. No doubt some of the English users in the present analysis were bilingual, perhaps even natives of the relatively collectivist Arab world (Hofstede, 2001). However, assuming that one is bilingual, the choice to tweet exclusively in English would generally suggest a certain affinity with the English-speaking (highly individualist) world. This limitation aside, in both analysis one and two, the hypothesised patterns of cultural influence were discernable and convergent with previous paper-based cross-cultural explorations (Dhawan et al., 1995; Ma and Schoeneman, 1997; Watkins et al., 2003). Confirmatory patterns were also observed for the individual-self primacy hypothesis (Gaertner et al., 1999), where personal attributes outnumbered social attributes across Twitter bios irrespective of language (Cousins, 1989; del Prado et al., 2007; Rhee et al., 1995). Similar findings, congruent with past research, were also observed for gender differences (Jackson et al., 1994).

The present study raises the prospect of using big data to explore self-concept and relative levels of idiocentric/allocentric values among individuals within a single national or cultural group. Furthermore, the use of social media data, which is continuously growing, facilitates surveillance and the exploration of self-concept across time, both at the individual and societal level.

Beyond finding support for the idea of cultural influences on self-construal, the present study also suggests that the systematic analysis of large social media datasets can be a useful adjunct to the traditional methods utilised in self-concept research. While these big data methods lack the control of experimental designs and the investigator defined structure of questionnaire-based surveys, they compensate for these shortcomings with volume (large amounts of data) velocity (exponential data growth) and variety (audio, video, text).

The present study has all of the advantages and limitations inherent in utilising big data. Furthermore, the use of social media data, specifically Twitter, limits generalizability to the broader population. Additionally, the present study only looked at the Twitter data for the UAE across 2015. Future studies might use global or multinational data sets and implement longitudinal or even prospective analyses, noting when and how people update their biographical information. With advances in data science providing increasingly sophisticated analytic tools, the exploration of self-concept through social media datasets seems a feasible and useful future direction.

Declarations

Author contribution statement

Justin Thomas, Aamna AlShehhi, Ian Grey, Marwa Al-Ameri: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interest statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

References

Baumeister, R.F., 1998. The self. In: Gilbert, D.T., Fiske, S.F., Lindzey, G. (Eds.), The Handbook of Social Psychology. Oxford University Press, New York, pp. 680–740.

Bond, M.H., Cheung, T.-S., 1983. College students' spontaneous self-concept. J. Cross Cult. Psychol. 14 (2), 153–171.

Cousins, S.D., 1989. Culture and self-perception in Japan and the United States. J. Personal. Soc. Psychol. 56 (1), 124–131.

del Prado, A.M., Church, A.T., Katigbak, M.S., Miramontes, L.G., Whitty, M., Curtis, G.J., et al., 2007. Culture, method, and the content of self-concepts: testing trait, individual-self-primacy, and cultural psychology perspectives. J. Res. Personal. 41 (6), 1119–1160.

Dhawan, N., Roseman, I.J., Naidu, R.K., Thapa, K., Rettek, S.I., 1995. Self-concepts across two cultures. J. Cross Cult. Psychol. 26 (6), 606–621.

Dodds, P.S., Harris, K.D., Kloumann, I.M., Bliss, C.A., Danforth, C.M., 2011. Temporal patterns of happiness and information in a global social network: hedonometrics and twitter. PLoS One 6 (12), e26752, 1–26.

Gaertner, L., Sedikides, C., Graetz, K., 1999. In search of self-definition: motivational primacy of the individual self, motivational primacy of the collective self, or contextual primacy? J. Pers. Soc. Psychol. Compass 76, 5–18.

Garcia-Gavilanes, R., Quercia, D., Jaimes, J., 2013. Cultural dimensions in twitter: time, individualism and power. In: Paper Presented at the Seventh International AAAI Conference on Weblogs and Social Media, Boston, Massachusetts USA.

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L., 2009. Detecting influenza epidemics using search engine query data. Nature 457 (19).

Heine, S.J., 2001. Self as cultural product: an examination of East Asian and North American selves. J Pers 69 (6), 881–906.

Hofstede, G., 2001. Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations across Nations, second ed. Sage Publications, London.

Jackson, L.A., Hodge, C.N., Ingram, J.M., 1994. Gender and self-concept: a reexamination of stereotypic differences and the role of gender attitudes. Sex. Roles 30 (9), 615–630.

Kanagawa, C., Cross, S.E., Markus, H.R., 2001. “Who Am I?” The cultural psychology of the conceptual self. Personal. Soc. Psychol. Bull. 27 (1), 90–103.

Kuhn, M.H., McPartland, T.S., 1954. An empirical investigation of self-attitudes. Am. Sociol. Rev. 19 (1), 68–76.

Ma, V., Schoeneman, T.J., 1997. Individualism versus collectivism: a comparison of Kenyan and American self-concepts. Basic Appl. Soc. Psychol. 19 (2), 261–273.

Markus, H.R., Kitayama, S., 1991. Culture and the self: implications for cognition, emotion, and motivation. Psychol. Rev. 98 (2), 224–253.

Rhee, E., Uleman, J.S., Lee, H.K., Roman, R.J., 1995. Spontaneous self-descriptions and ethnic identities in individualistic and collectivistic cultures. J. Personal. Soc. Psychol. 69 (1), 142–152.

Thomas, J., 2014. Psychological Well-Being in the Gulf States: the New Arabia Felix. Palgrave Macmillan, London.

Watkins, D., Adair, J., Akande, A., Gerong, A., McInerney, D., Sunar, D., 1998. Individualism-collectivism, gender, and the self-concept: a nine culture investigation. Psychologia 41, 259–271.

Watkins, D., Cheng, C., Mpofu, E., Olowu, S., Singh-Sengupta, S., Regmi, M., 2003. Gender differences in self-construal: how generalizable are Western findings? J. Soc. Psychol. 143 (4), 501–519.

Yang, C., Srinivasan, P., 2016. Life satisfaction and pursuit of happiness on twitter. PLoS One 11 (3), 1–30.