Machine Learning with Applications 10 (2022) 100419

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Who believes they are good navigators? A machine learning pipeline highlights the impact of gender, commuting time, and education

You Cheng [a,*,1], Chuanxiuyue He [b,*,1], Mary Hegarty [b], Elizabeth R. Chrastil [a,c]

[a] Department of Cognitive Sciences, University of California Irvine, Irvine, 92697, CA, USA
[b] Department of Psychological and Brain Sciences, University of California Santa Barbara, Santa Barbara, 93106, CA, USA
[c] Department of Neurobiology and Behavior, University of California Irvine, Irvine, 92697, CA, USA

ARTICLE INFO

Dataset link: https://github.com/LilianYou/Sea_Hero_Quest

Keywords:
Reproducibility crisis
Machine learning
Big data
Self-reports
Demographics
Spatial navigation

ABSTRACT

Large scale digital data, which are becoming more prevalent, offer the potential to alleviate reproducibility concerns in psychology research findings. However, large scale digital data are not sufficient in and of themselves, thus necessitating the need for the development of machine learning (ML) pipelines that are capable of handling high dimensional datasets at scale. Such ML-based methodologies enable the analysis of complex relationships, which allows for the consideration of complicated demographics, a factor that is likely to play a role in the generalizability of research. We introduce a novel ML pipeline and demonstrate its potential on a large-scale digital dataset, Sea Hero Quest, a mobile game with data from nearly 770,000 players (ages 19 to 70, men N = 404,455, women N = 367,173). We analyzed how demographics are related to self-reported navigation ability using exploratory analysis, supervised and unsupervised learning. The results suggest that gender is the most important demographic factor in predicting self-reported navigation ability, followed by daily commuting time, age, and education, such that men (compared to women), long commuters (compared to those whose commuting time is shorter than 1 h), and older people with tertiary education (compared to younger people with secondary education) tended to evaluate themselves as better navigators. The large-scale dataset and ML pipeline capture influential factors, such as daily commuting time and education level, which have often been overlooked and are difficult to investigate with in-laboratory studies that use limited samples and traditional analytical techniques.

1. Introduction

1.1. The reproducibility crisis in psychology

Over the past decade, one critical question in the field of psychology research is the reproducibility of research findings (aka. the reproducibility crisis), such that findings reported in one study cannot be replicated using an independent group of subjects (Camerer et al., 2018; Simmons, Nelson, & Simonsohn, 2011). Various methods have been proposed to solve the reproducibility crisis, including conducting restrained experimental designs (Ioannidis, 2005; Simmons et al., 2011), generating and sharing registered reports (Van't Veer & Giner-Sorolla, 2016), and archiving unpublished data in open databases (Schooler, 2011). However, a concern is that these methods simply narrow down the generalizability of research findings to a particular group of people.

Subjects in most psychology studies have traditionally been drawn from small and specialized samples — usually composed of college students in their early 20s. Moreover, sample source itself could be a possible explanation for failure to reproduce psychological findings. This is because samples may differ in their behaviors due to different demographic backgrounds (e.g., age, gender, education levels), which limits the generalizability of research findings. This calls for data at scale with well-representative samples that represent the diverse demographics of the human population (Ioannidis, 2005; Simmons et al., 2011).

1.2. Utilizing large-scale digital data as a solution to the reproducibility crisis

Digital data regarding people's behavior collected online through platforms such as Amazon Mechanical Turk, social media, or phone-based games are becoming more and more prevalent in the field of psychology research, especially during the pandemic when conducting in-person experiments has been challenging. Digital data usually have very large sample sizes that could result in thousands or even millions of data points, and also enable us to more easily sample from diverse

* Corresponding authors.
E-mail addresses: youc3@uci.edu (Y. Cheng), c_he@ucsb.edu (C. He), hegarty@ucsb.edu (M. Hegarty), chrastil@uci.edu (E.R. Chrastil).
[1] Cheng and He are the co-first authors on this paper.

https://doi.org/10.1016/j.mlwa.2022.100419
Received 15 July 2022; Received in revised form 2 October 2022; Accepted 2 October 2022
Available online 5 October 2022

Table 1
Summary table of the related work.

Related work	Research topic	Evidence for	Sample size	Limitation
Berteau-Pavy, Park, and Raber (2007), Coughlan, Laczó, Hort, Minihane, and Hornberger (2018), Kunz et al. (2015) and Puthusseryppady, Morrissey, Spiers, Patel, and Hornberger (2022)	Navigation Ability & Alzheimer's Disease	The potential of navigation ability as a cognitive fingerprint to detect incipient Alzheimer's disease	N = 15–115	Small sample size Limited behavioral measures
Li and Klippel (2016), Montello (2005), Nazareth, Huang, Voyer, and Newcombe (2019), Pagkratidou, Galati, and Avraamides (2020), Weisberg and Newcombe (2018) and Wolbers and Hegarty (2010)	Multifaceted Navigation Ability	Different tasks measure different aspects of navigation ability	Various across 20–100	Small sample sizes Lack of standard comprehensive measurement
Donald Heth, Cornell, and Flood (2002), Epstein, Higgins, and Thompson-Schill (2005), Hegarty, Montello, Richardson, Ishikawa, and Lovelace (2006), Hegarty, Richardson, Montello, Lovelace, and Subbiah (2002), Hund and Padgitt (2010), Meneghetti, Borella, Pastore, and De Beni (2014), Pazzaglia, Meneghetti, Labate, and Ronconi (2016) and van der Ham, van der Kuil, and Claessen (2021)	Self-reported Navigation Ability	Small to moderate associations with performance in objective navigation behavior tasks	Various across 20–7150	Limited demographics information Limited connections to behaviors
Lester, Moffat, Wiener, Barnes, and Wolbers (2017), Nazareth et al. (2019) and Spiers, Coutrot, and Hornberger (2021)	Group differences in Navigation	Robust gender differences and aging deficits in navigation ability	Various across 20–250	Limited sample sizes Limited demographics analysis
Coughlan et al. (2019), Coutrot et al. (2022, 2018) and Spiers et al. (2021)	Sea Hero Quest: Population level navigation ability	A population-level benchmark performance in terms of mobile-based navigation tasks	Over 500,000	Limited self-reports and demographics analysis

This table lists important studies on individual differences in navigation ability and illustrates limitations of these previous studies, including relatively small sample sizes, inconsistent measures, and insufficient demographic reports.

populations (e.g., in terms of age, gender, race/ethnicity, education, socioeconomic status, handedness, etc.).

1.3. Incorporating demographic information in large scale digital data to systematically interpret self-report measures

Self-report is a common method in psychology research in which people are asked to directly report their feelings, attitudes, beliefs, or behaviors (Jupp, 2006). It can be carried out in multiple forms, including open and closed-ended questions, rating scales, interviews, etc. Self-report is more efficient than physiological and behavioral measurements, which are expensive in terms of both time and labor cost. It is an economic approach to prescreening targeted samples and in prognosing neurological disorders (e.g., Blazer, Hays, Fillenbaum, & Gold, 1997; Jonker, Launer, Hooijer, & Lindeboom, 1996; Taylor, Miller, & Tinklenberg, 1992). However, self-report measurements have been susceptible to concerns about validity – the extent to which a measure is indeed measuring what it claims to measure – as respondents may underestimate or overestimate their behaviors. Such mischaracterization could be intentional (Jupp, 2006) or could emerge from subconscious social stereotypes related to their demographic backgrounds (Reychav et al., 2019; Slavin et al., 2010; van der Ham et al., 2021; Wasef et al., 2021). In the latter situation, the reproducibility and generalizability of previous findings based on small samples can be examined using large datasets in which respondents represent a spectrum of demographic characteristics. This raises a new question as to how to thoroughly analyze the relationship between demographic information and human behaviors in large scale digital datasets.

Here, we propose a machine learning pipeline for analyzing demographic relevance in large-scale digital self-report data. We tested the pipeline with a large dataset (around 770,000 global users' data) collected from a phone-based game – Sea Hero Quest – in which people play navigation games in virtual environments, report their demographics, and evaluate their own navigation abilities (Coutrot et al., 2022, 2018; Spiers et al., 2021). One goal of Sea Hero Quest is to set a population-level benchmark for dementia research. In the study, people's self-evaluations of their navigation abilities are based on a Likert-scale measurement, which is one of the most common types of rating scales in self-report measurement. Further, this ML pipeline could easily be generalized to detect behavior patterns in large scale self-report data in a wide variety of psychology research studies in the future.

2. Related work

Spatial navigation is a critical cognitive ability which enables people to represent their environments so as to reach target locations efficiently without getting lost or experiencing anxiety. Previous literature has emphasized the potential of navigation ability as a cognitive fingerprint to detect incipient Alzheimer's disease (Berteau-Pavy et al., 2007; Coughlan et al., 2019, 2018; Kunz et al., 2015; Puthusseryppady et al., 2022). Nevertheless, we still do not have a standard comprehensive task battery to measure individual navigation ability. Table 1 summarizes previous literature showing (1) the relations between navigation ability and Alzhemier's disease; (2) self-reports as a promising efficient measure given the limitations of the objective measures; (3) previous research on demographics and navigation ability; and (4) the Sea Hero Quest project, which took the initiative to develop a population benchmark of the navigation ability. In the table, we also emphasize the limitations of the current approaches to each research topic. Specifically, objective measures of navigation ability are difficult to acquire on a large scale. In contrast, self-report measures are easier

to acquire, but may be less reliable. Further, the relationship between demographic information such as age, gender, etc. and both objective and self-report measures has only been studied with relatively small sample sizes. We elaborate on each of these points and the details of the table in the following sections.

2.1. The lack of efficient measures to study spatial navigation ability

Navigation ability is usually measured using various paradigms or metrics, such as the ability to memorize landmarks, learn routes, form an accurate and coherent mental representation of the whole environment (i.e., cognitive map), estimate directions and distances, and give efficient directions (Hegarty, Burte, & Boone, 2018; Montello, 2005; Weisberg & Newcombe, 2018; Wolbers & Hegarty, 2010). People's performance also varies in different environments and with different task goals (Li & Klippel, 2016; Nazareth et al., 2019; Pagkratidou et al., 2020). It is time consuming and labor intensive to collect valid data to evaluate healthy participants' navigation ability on a large enough scale that could illustrate the distribution of abilities across age and other demographic information (e.g., home environments, commuting time, etc.).

In contrast, self-reported navigation, as an easier and more economical measure, has shown small to moderate associations with performance in objective navigation tasks (Donald Heth et al., 2002; Epstein et al., 2005; Hegarty et al., 2002; Hund & Padgitt, 2010; Meneghetti et al., 2014; Pazzaglia et al., 2016), demonstrating its potential to be used as a powerful prescreening tool for detecting neurological disorders that have navigation impairments. However, the relationship between self-reported navigation ability and demographic information is unclear, which calls for more research on generalizing findings based on small samples to the general public with various demographic backgrounds.

2.2. Demographics and navigation ability

Previous studies have found solid evidence supporting gender differences and aging deficits in navigation ability at a behavioral level (Lester et al., 2017; Nazareth et al., 2019; Spiers et al., 2021). Gender and age effects have been discussed in self-reported measures as well (e.g., Hegarty et al., 2006; van der Ham et al., 2021). However, these studies used relatively small samples and as a consequence could not take advantage of modeling methods targeting a large dataset (e.g., clustering or random forest). Using a large dataset with a ML-based analyses pipeline enables us to investigate whether people's self-evaluations match the findings of those in-laboratory empirical studies. Preliminary evidence based on over 7000 participants in an online study showed that older men tended to overestimate their navigation ability as measured in online video-based navigation tasks (van der Ham et al., 2021). Although that study captured the influence of gender and age on self-reported navigation abilities reasonably well, we propose here that large-scale digital data could advance our knowledge further by considering other demographic information (e.g., education levels, home environments, etc.) with machine learning tools.

The Sea Hero Quest dataset (SHQ) is composed of both large-scale self-reported navigation ability data and multidimensional demographic data (more details below) in addition to data from multi-level objective navigation task performance (which will not be considered in this report). Therefore, the large SHQ dataset enables us to test for relationships between people's self-reported navigation ability and their demographic information more systematically than previous studies. More importantly, we demonstrated our ML-pipeline via the analyses, which could be generalized to analyze other large-scale digital data in future psychology research.

3. Methodology

The first step in our ML-pipeline[2] was to conduct a correlation analysis to explore the relationships between the observed variables in the Sea Hero Quest sample. These were age, gender, handedness, education level, home environment (i.e., rural (level 1), city (level-3) or in-between/mixed (level 2)), average daily sleep hours, average daily commute time, and self-reported navigation ability. Second, we performed factor analysis to detect potential latent variables underlying our observed variables. Third, we implemented an unsupervised method (k-means clustering) to detect subpopulations in the sample, based on the demographic information. Fourth, we implemented a chi-squared independence test, which allowed us to determine how self-reported navigation ability varies at each rating level, and across detected subpopulations. These relationships were used to form data-driven theories. Fifth, we used a supervised learning method (ordinal logistic regression) to detect relationships between demographics and self-reported navigation ability. Lastly, we implemented another supervised learning method, complementing our parametric model with a non-parametric model (random forest regression), which yielded consistent results with our regression analysis. See Fig. 1.

We conducted our analyses on Google's cloud server using Colab notebooks with 2.3 GHz CPU Frequency, 2 CPU Cores (Haswell), and 12 GB RAM.

3.1. Data description

We started with the full raw dataset which included approximately 4 million users (Coughlan et al., 2019). Because demographic and self-reported navigation ability questions were all optional in the game, not every user answered all questions. Accordingly, we only included users who responded to all questions (except for the countries, which is out of the scope of this study, see more in Coutrot et al., 2018). Then, based on previous research, we excluded people who reported sleeping less than 3 h or over 12 h[3] every day, reported being younger than 19 years old or over 70 years old,[4] did not identify as a male or female, or reported to have "unspecified" education (see Fig. 2). Our following analyses are based on 771,628 users (52.4% male).

Fig. 3 illustrates the distributions of self-reported navigation ability and of all demographic variables in the sample. Around 90% of participants were right-handed and the gender ratio was around 0.5, which is representative of the world population. Most of the participants were in their early twenties and 71.8% participants had a tertiary level of education (i.e., college or university), indicating that our sample was relatively younger and had higher education levels than the world population. Note that most participants reported "good" for self-reported navigation ability (54.5%) and that only 13.6% participants reported being "bad" or "very bad" at navigation.

3.2. Exploratory data analyses

3.2.1. Correlation analysis

We first looked at bivariate Spearman correlations among all variables (see Fig. 4). Correlations among most variables were lower than .1, suggesting a low probability of latent factors (See more in the Factor Analyses section). However, age and sleep, age and daily commuting

[2] It is worth noting that this is not a fully automatic ML-pipeline, but rather a sequence of ML analytics.

[3] Normal light sleepers still sleep over 3 h per day. People who reported sleeping over 9 h per day are considered as long sleepers and rarely will people sleep over 12 h every day (Grandner & Drummond, 2007; Patel, Malhotra, White, Gottlieb, & Hu, 2006). Thus, we filtered out data by people who reported sleeping less than 3 h or over 12 h.

[4] Based on previous research on Sea Hero Quest (Coughlan et al., 2018), people who reported being younger than 19 years old or older than 70 years old showed abnormal behaviors.

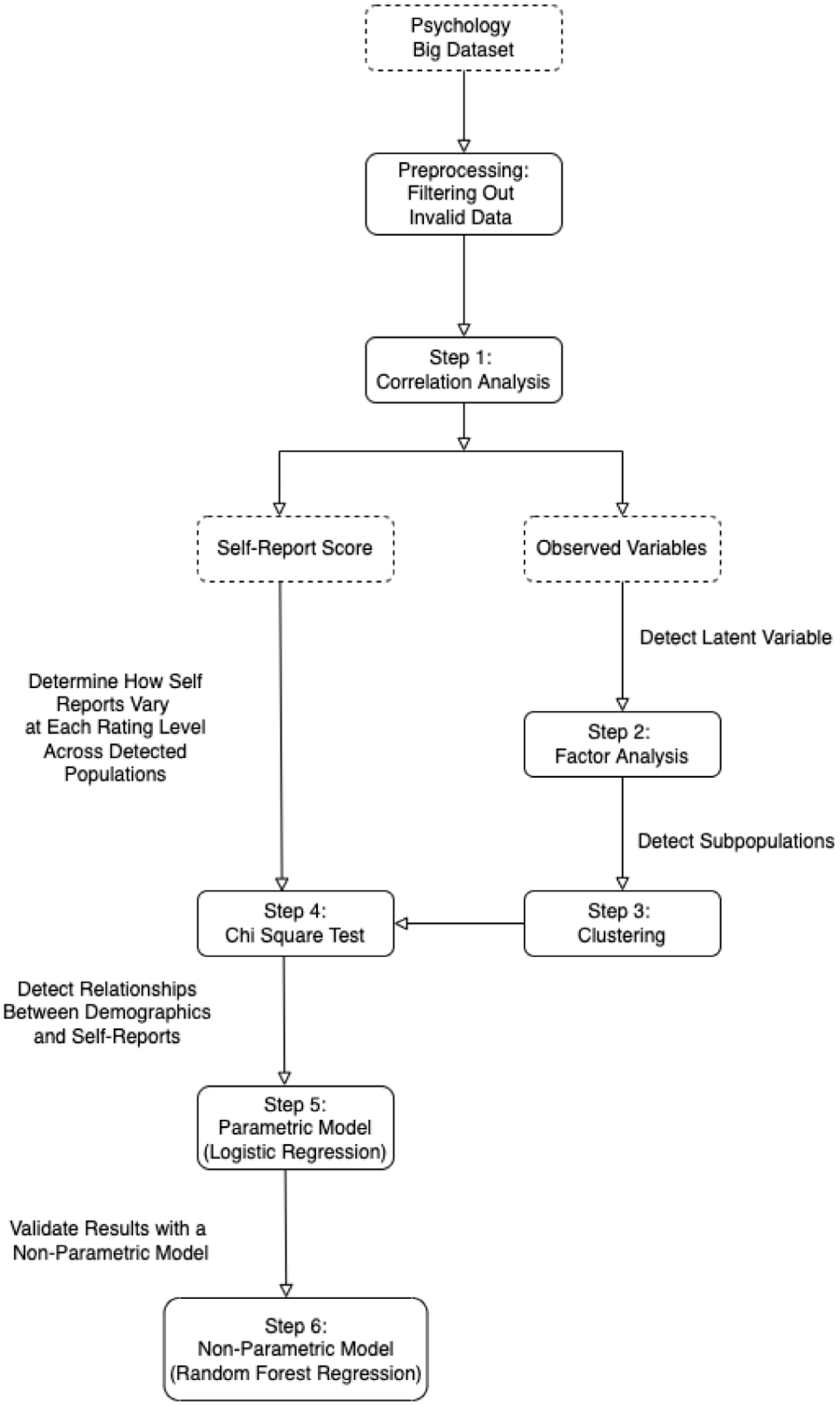

Fig. 1. The proposed pipeline. *Note:* Dashed boxes indicate data modules, solid boxes indicate processing modules.

time, as well as gender and commuting time were significantly correlated with each other (rs are above .1, ps $< .001$) and male participants tended to report better navigation ability ($r = .24$, $p < .001$). With a large sample size, even small correlations will be significant; therefore, we did not consider effect sizes below 0.1, which were judged to be too small to be meaningful.

3.2.2. Factor analysis

We first conducted a Bartlett Sphericity Chi Square Test on all independent variables to test whether there was a pattern among the independent variables. The test result was significant ($Score = 109\,858$, $p < .001$), indicating that there was such a pattern. Note that in this and all following analyses, we did not include self-reported navigation

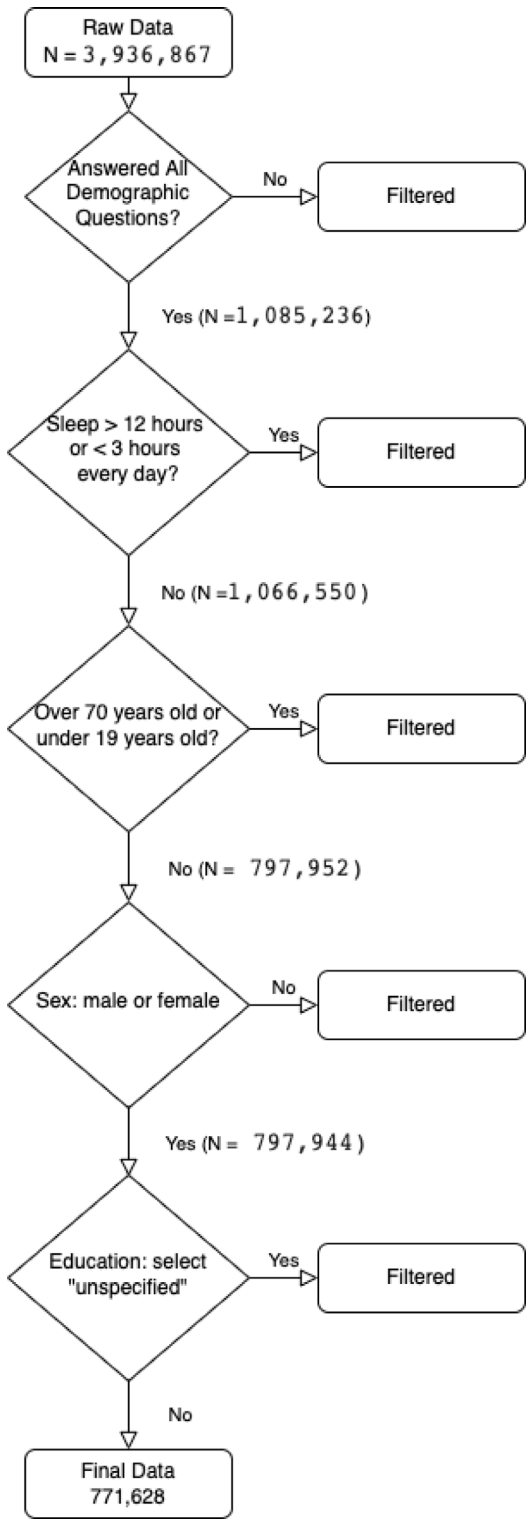

Fig. 2. Flowchart of the data filtering for the Sea Hero Quest self-reported database.

ability as a factor, it was only part of the analysis when it was used as a dependent variable in logistic regression and random forest regression.

Next, we conducted a Kaiser–Mayer–Olkin (KMO) Test to test whether there was sufficient variance in the dataset to conduct a factor analysis. The KMO score was .508, which is smaller than the criterion of .6 (Kaiser, 1974), indicating there was not sufficient variance for factor analyses.

Thus, although the Bartlett Sphericity Chi Square Test showed that there was a pattern among the independent variables, there was no latent factor. This finding suggests patterns emerge from the independent variables alone, which can be tested by cluster analyses.

3.2.3. Clustering

Subpopulations. To determine whether there were subpopulations in the dataset, we conducted k-means clustering analyses, based on all of the demographic variables except self-reported ability. K-means clustering – partitioning observations into k clusters where each observation is assigned to one cluster with the nearest mean – is one of the simplest and most computationally efficient partitioning methods (Forgy, 1965; Lloyd, 1982). It has been commonly used for partitioning people into subpopulations based on their demographics (e.g., customer segmentation in marketing to construct customer profiles) in many industries (e.g., Kansal, Bahuguna, Singh, & Choudhury, 2018; Namvar, Gholamian, & KhakAbi, 2010; Wu, Yau, Ong, & Chong, 2021).

The k-means clustering was conducted using the Python sklearn package. Because k-means clustering uses the Euclidean distance for measuring object similarities, all variables were first preprocessed by normalizing to the range between 0 and 1. The initial 4 cluster centroids were selected randomly from the data. The clustering analyses yielded 4 clusters based on the elbow method, which were then validated with additional methods such as Davies Bouldin score and Silhouette score to identify the optimal number of clusters. The model took 2.11 s to run (CPU times: user 2.1 s, system: 27.3 ms).

As shown in Fig. 5, the four clusters (called groups) represent four subpopulation groups. Group 1 (called "male long commuter") was composed of males with education levels close to that in the total sample distribution, and a daily commute of more than 1 h. Group 2 (called "female tertiary education") was composed of females with tertiary education, and a daily commute close to that in the total sample distribution. Group 3 (called "male tertiary education short commuter") was composed of males with tertiary education, and a daily commute of less than 1 h. Group 4 (called "secondary education") was composed of people with secondary education, equally representative of both genders and with a daily commuting time close to that of the total sample distribution. Further, age, sleep, home environment, and handedness were evenly distributed across the four groups. These four subpopulation groups differ in their demographics, especially in terms of gender, education, and daily commuting time. Thus, we focus on the interactions between these variables for the ordinal regression analysis.

Subpopulations and Self-Reported Navigation Ability. We then tested whether self-reported navigation ability varies for people in the different subpopulations. First, the average self-reported navigation ability of each group was significantly different from each of the others (Non-parametric ANOVA with Conover's post hoc test and Holm–Bonferroni Correction, $p < 0.001$). More specifically, male long commuters reported the best navigation skills, followed by male short commuters with tertiary education, people with secondary education, and lastly females with tertiary education (See Fig. 6).

Then, we conducted the chi-square test of independence to test the frequency distribution of all levels of self-reported navigation skills among the four subpopulations. This analysis showed that group membership and self-reported navigation skills were significantly associated ($x^2 = 41069.51, p < 0.001$). Post-hoc pairwise comparisons revealed that all groups at all navigation ability levels were significantly different from each other (all $p < 6.25e{-}5$, Bonferroni corrected). In other words, the percentages of people that indicated that their navigation ability as "very bad", "bad", "good", or "very good" were different across groups (See Fig. 7). More specifically, male long commuters reported the highest within-group percentage of "very good" at navigation, followed by male short commuters with tertiary education, people with secondary education, females with tertiary education. People with secondary education reported the highest within-group percentage of

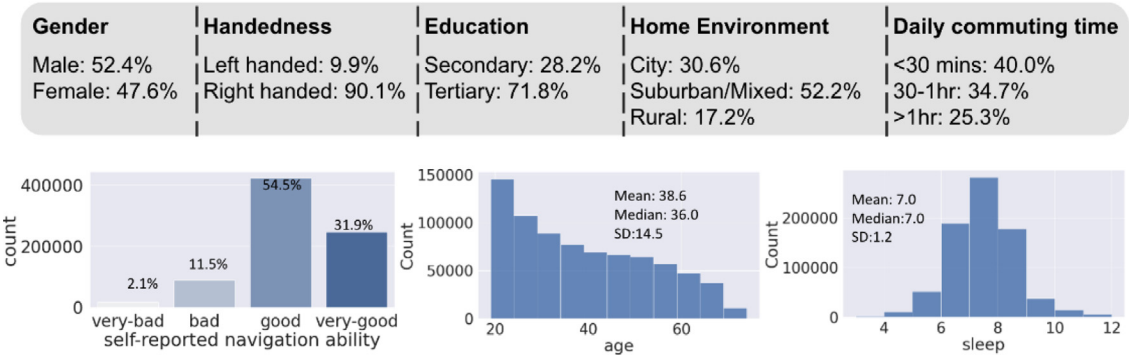

Fig. 3. Percentages and histograms of all self-reported variables.

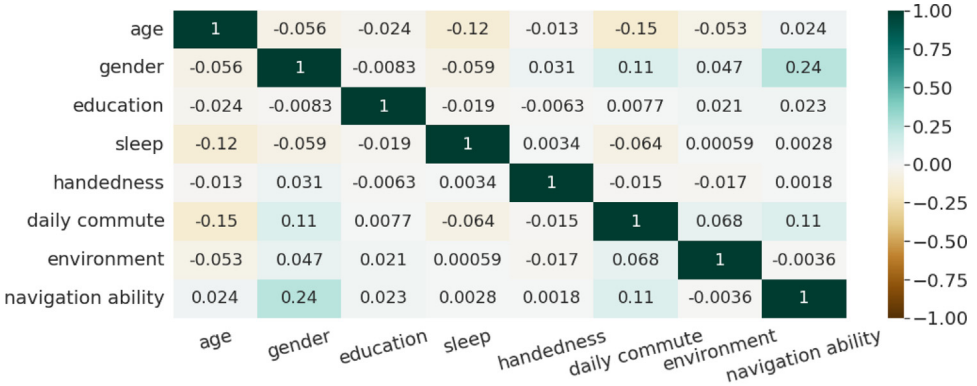

Fig. 4. Bivariate spearman correlations between the self-reported variables. *Note:* All correlations were statistically significant ($p < 0.001$).

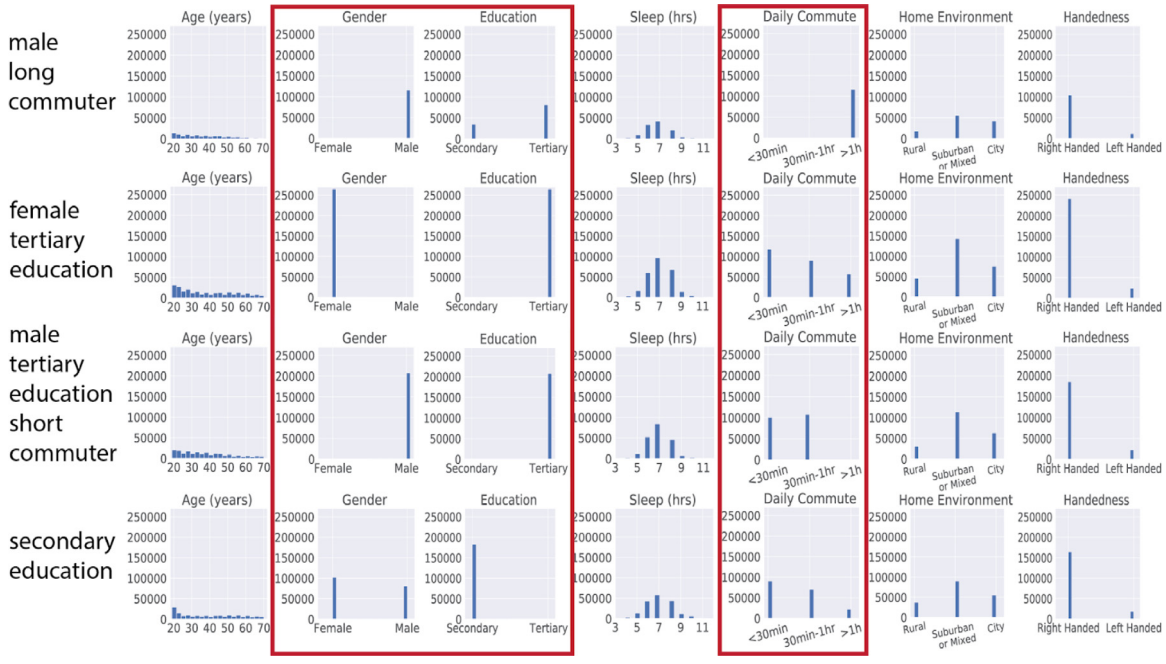

Fig. 5. Subpopulations featured by histograms of seven demographic factors. Male long commuter: Males across both education levels with more than 1 h daily commute. Female tertiary education: Females with tertiary education that have a wide range of daily commuting times. Males tertiary education short commuter: Males with tertiary education that commute less than 1 h on a daily basis. Secondary education: Both males and females with secondary education that have a wide range of daily commuting times. The differences among four groups were mainly driven by gender, education, and commuting time. Age, sleep, home environment, and handedness were evenly distributed for each group.

"good" at navigation, followed by females with tertiary education, male short commuters with tertiary education, male long commuters. Females with tertiary education reported the highest within-group percentages of both "bad" and "very bad" at navigation, both followed by

Table 2
Coefficients of the ordinal linear regression model.

Factors	Coeff	Effect size	z-value	*p*-value
age	0.058	1.06	18.30	<.001
gender	0.851	2.34	100.12	<.001
age:gender	0.071	1.07	15.62	<.001
edu	0.043	1.04	6.11	<.001
gender:edu	0.138	1.15	13.873	<.001
commute	0.324	1.38	66.72	<.001
gender:commute	−0.082	0.92	−17.98	<.001
edu:commute	−0.137	0.87	−27.98	<.001
left-hand	−0.024	0.98	24.11	=.001
sleep	0.055	1.06	−17.54	<.001
city-like	0.32	0.96	0.17	<.001

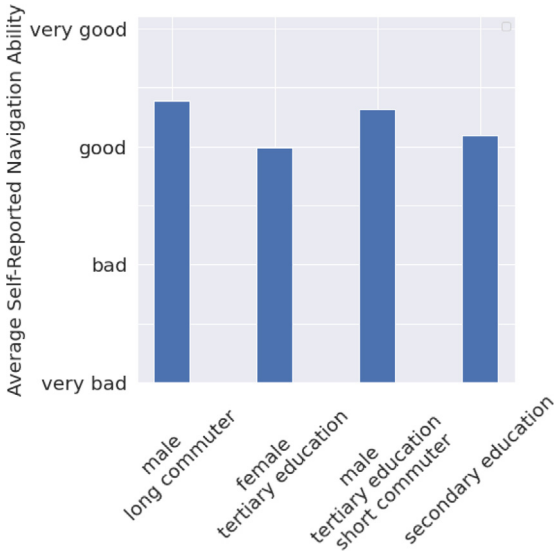

Fig. 6. Average navigation ability in each subpopulation group. Male long commuter: Males across both education levels with more than 1 h daily commute. Female tertiary education: Females with tertiary education that have a wide range of daily commuting times. Males tertiary education short commuter: Males with tertiary education that commute less than 1 h on a daily basis. Secondary education: Both males and females with secondary education that have a wide range of daily commuting times. The differences among four groups were mainly driven by gender, education, and commuting time. Age, sleep, home environment, and handedness were evenly distributed across the four groups. *Note:* Standard error was too small due to large sample size to be visible on these graphs.

people with secondary education, male short commuters with tertiary education, male long commuters.

Next, we looked into how factors, especially those driving the clusters (i.e., gender, education, and commuting time), relate to self-reported navigation ability. Specifically, we conducted a follow-up analysis, reported in the next Section 3.3 to test the importance of these factors, and their interactions, in predicting self-reported navigation ability.

3.3. Ordinal logistic regression

Ordinal logistic regression was utilized here given that (1) the target variable is in an ordinal scale (2) the regression analysis is computationally economical and (3) interpretable. It is a good starting point to examine the relationship between the predictor variables and the target variable. We utilized the statsmodel package in Python to construct the ordinal model. The logit link function (distribution) was used and the fitting method is Broyden–Fletcher–Goldfarb–Shanno (BFGS) which is broadly used for fitting logistic regression (Kochenderfer & Wheeler, 2019).

Previous literature suggested that there could be interactions of age by gender (van der Ham et al., 2021; Yu et al., 2021). Thus, these

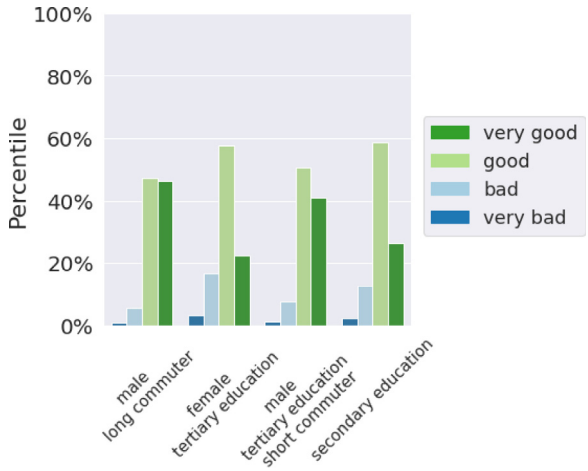

Fig. 7. Percentage of self-reported navigation ability levels in each subpopulation group. Groups and levels of self-reported navigating skills were significantly related (x2 = 41 069.51, p < 0.001). All groups at all navigation ability levels were significantly different (all p < 6.25e−5). Bonferroni corrected.

interaction variables, as well as the primary variables of age, gender, education level, daily commuting time, daily sleep hours, and home environment were added to the ordinal logistic regression model. Age, daily commuting time and sleep hours were standardized and centered. The whole model took 9 min 57 s to run (CPU times: user 6 min 24 s, system: 3 min 33 s).

The results showed that the factors (including the interactions) all significantly contributed to the model ($X^2(14) = 55\,302.18$, $p < .001$, pseudo $R^2 = .035$). The effect sizes in Table 2 can be interpreted as an odds ratio, which measures how many times the odds of reporting good navigation ability increases if the factor increases one standard deviation (for continuous variables) or changes to another value (for categorical variables). If this value is 1, it means the odds do not change. For example, the odds of males reporting higher navigation ability was 2.34 times that of females, which is statistically significant, $z(771\,614) = 100.12$, $p < .001$. We did not consider effect sizes between 0.9–1.1, which were judged to be too small to be meaningful, even if they were significant, as the significance was likely due to the large sample size. Thus, we focused on the factors with relatively large effect sizes and left the other interesting trends for future studies.

To sum up, based on the ordinal logistic regression, men who commuted longer every day had higher odds of reporting better navigation ability, see Table 2. In terms of the interactions or the moderators, the strongest effect was education by gender which suggested that the gender gap was stronger for people with tertiary education than people with secondary education. The second strongest effect was education by commuting time, which implied that the commuting time effect was stronger for people with tertiary education than for people with secondary education.

3.4. Random forest regression

To complement our parametric model of ordinal logistic regression, we also conducted a non-parametric model to examine the importance of demographic variables in predicting self-reported navigating skills. We applied the random forest method, which is a meta estimator that utilizes ensembled decision trees (Breiman, 2001). The random forest method has high computational efficiency because its child estimator (i.e., standard tree growing algorithms) has low computational cost; the method also prevents overfitting by using multiple trees. We chose the random forest regression model rather than the classification model because the former could incorporate ordinal information in the dependent variable (Janitza, Tutz, & Boulesteix, 2016). We predicted

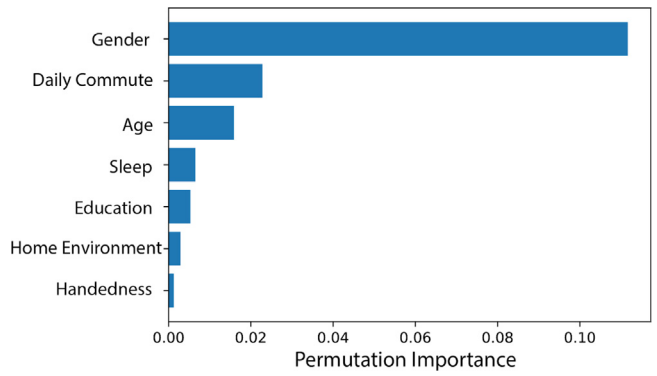

Fig. 8. Permutation based feature importance ranking in the random forest regression analysis.

self-reported navigation ability based on gender, age, sleep, education, home environment, daily commute, and handedness.

In the analysis, data were randomly split into a 75% training set and a 25% testing set. The RandomForestRegressor function in the Python sklearn package was used. We used 100 trees, each tree was built on bootstrapped samples given equal weight, with the quality of each split measured based on mean squared error (i.e., variance reduction), and the node size set to default (i.e., expanded until all leaves are pure). Features are always randomized at each split and all features are considered to split a node (i.e., bagged trees) as empirically justified in Geurts, Ernst, and Wehenkel (2006). For the variable importance measurement, we used permutation-based importance ranking as it gives more unbiased rankings of the predictors (Altmann, Toloşi, Sander, & Lengauer, 2010). Permutation importance was computed based on the held-out test set. The model took 1 min 15 s to run (CPU times: 1 min 14 s, system: 149 ms).

Permutation based feature importance rankings from the random forest analyses revealed that gender was the most important variable for predicting self-reported navigating skills, followed by commuting time, age, average hours of sleep, education levels, home environments, and handedness (See Fig. 8). This permutation importance ranking aligns with the results of the logistic regression, further supporting the order of different demographic variables in predicting self-reported navigation ability.

4. Conclusions

4.1. Summary of findings

A machine-learning based analysis sequence was used to investigate multidimensional demographic information in a large-scale digital dataset. The pipeline used a combination of descriptive analyses, exploratory analyses, parametric supervised learning models (ordinal regression), non-parametric unsupervised learning models (clustering methods), and non-parametric supervised learning models (random forest). The pipeline was tested with data from the mobile game Sea Hero Quest with approximately 770,000 users. The results identified gender as the most important demographic factor in predicting self-reported navigation ability, followed by daily commuting time. The clustering, regression, and random forest models also identified age, education, and daily sleep hours as important factors.

4.2. Theoretical contributions

This analysis provides three theoretical contributions to the spatial navigation domain. First, we found a gender effect on the self-report measure at a large scale. This result is consistent with findings from previous studies with relatively small samples regarding individual differences in spatial navigation (Hegarty et al., 2022; Montello, 2005; Weisberg & Newcombe, 2018; Wolbers & Hegarty, 2010).

Second, these analyses provide a new interpretation of the effect of age on navigation ability. In previous studies, researchers found that people's navigation performance decreased with age (Coutrot et al., 2018; Lester et al., 2017; Yu et al., 2021) but their self-reported navigation ability increased with age (van der Ham et al., 2021). A few of these studies showed that the age effect was moderated by gender (Coutrot et al., 2018; van der Ham et al., 2021; Yu et al., 2021). Our results also replicated that the age effect on self-reported navigation ability was moderated by gender, however, the age effect was not as strong as that in previous research, nor was the interaction. This suggests interesting future directions for investigating the age effect.

Furthermore, this study highlighted the importance of daily commuting time and education level on self-reported navigation ability, even after adjusting for effects of gender and age. Specifically, we found that people who commuted longer on a daily basis and people who had tertiary education tended to report better navigation ability. Assuming there are associations between self-reported navigation ability and one's actual navigation performance, the commuting effect is consistent with the argument that navigation ability is a "use-it-or-lose-it" skill (McKinlay, 2016). The main effect of education was relatively small, but the interaction with gender was substantial. The education effect is in line with previous neuroscience literature that emphasizes the importance of education (as part of socioeconomic status) in brain development (Farah, 2017; Poeppl et al., 2022). It is clear that to better understand these interesting and novel mediation and moderation effects, further studies are required.

4.3. Future directions

Although the sample for Sea Hero Quest has covered people with much more variable demographics than that of typical psychology samples (often based on college students), our investigation has shown that it is still not representative of the world population. One limitation is that a large proportion of subjects were in their 20 s and had tertiary education. In addition, playing Sea Hero Quest requires people to have access to smartphones and the internet. The sample is likely not representative of the world population partially due to the fact that such digital devices are not accessible to everyone equally. To make the ML pipeline more generalizable, more representative population-level data would be required in future research. However, acquiring a large-scale digital dataset that is representative of the world population would be challenging for the reason suggested above. Indeed, these could be common issues for other large-scale digital data used in future research in psychology. Our results highlight the importance of analyzing confounds driven by the demographics of the participants in future large datasets analyses.

Second, although previous research has demonstrated the predictive power of self-reported navigation ability on objective navigation ability (Hegarty et al., 2002), these measures are far from being equivalent. Specifically, self-reported navigation ability is only moderately correlated with performance in navigation tasks, such as retracing a route taken previously, learning the layout of new places from different views and navigation experiences, and estimating directions and distances to known locations (Donald Heth et al., 2002; Epstein et al., 2005; Hund & Padgitt, 2010; Meneghetti et al., 2014; Pazzaglia et al., 2016). Thus, objective measurements are still necessary and can continue to help to evaluate the validity of self-reports. Comparing objective measures and self-reports with large samples such as this might help to identify potential subconscious biases in self-reports that are related to their demographic backgrounds. Further, if a link between objective measures and self-reports could be established, using self-reports could provide an efficient prognosis for people with neurological disorders that have navigation impairments.

Third, preliminary evidence has suggested that individual task performance in Sea Hero Quest is predictive of one's navigation ability in

real life (Coutrot et al., 2019), but more work needs to be done to paint a full picture, including participants' demographics.

In the future, pretrained models from the Sea Hero Quest dataset could be used to improve prediction accuracy for psychology studies with smaller samples. In that case, psychologists who study similar questions could more systematically approach their findings. This approach constitutes a useful addition to a growing set of methods which may collectively help alleviate the reproducibility crisis currently facing the field of experimental psychology.

4.4. Broader implications

By incorporating multivariate demographics in big data analyses, we demonstrated an approach that not only comprehensively interprets self-report results, but also informs the reproducibility of these results from a new perspective. Specifically, we replicated the findings in the experimental literature with new interpretations by incorporating new demographics at scale.

Specific to the research field of spatial navigation, individual differences in self-reported navigation ability have been linked to individual differences in Global Positioning System (GPS) use (He & Hegarty, 2020; Hejtmánek, Oravcová, Motýl, Horáček, & Fajnerová, 2018). Understanding more about user characteristics paves the way towards more efficient human–GPS interactions. In clinical settings, these analyses inform the future development of an adaptive self-reported threshold for preclinical screening based on demographic factors (Coughlan et al., 2019, 2018; Spiers et al., 2021).

CRediT authorship contribution statement

You Cheng: Methodology, Formal analysis, Conceptualization, Writing – original draft, Writing – review & editing. **Chuanxiuyue He:** Methodology, Formal analysis, Conceptualization, Writing – original draft, Writing – review & editing. **Mary Hegarty:** Funding acquisition, Writing – review & editing, Supervision. **Elizabeth R. Chrastil:** Funding acquisition, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

We do not have the permission to share the data but we have shared our code on Github https://github.com/LilianYou/Sea_Hero_Quest.

Acknowledgments

This project was supported by the National Science Foundation, USA under Grant No. NSF IIS-2024633. We thank Hugo J. Spiers for his suggestions and for making the dataset available to us.

References

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347. http://dx.doi.org/10.1093/bioinformatics/btq134.

Berteau-Pavy, F., Park, B., & Raber, J. (2007). Effects of sex and APOE ε4 on object recognition and spatial navigation in the elderly. *Neuroscience*, *147*(1), 6–17. http://dx.doi.org/10.1016/j.neuroscience.2007.03.005.

Blazer, D. G., Hays, J. C., Fillenbaum, G. G., & Gold, D. T. (1997). Memory complaint as a predictor of cognitive decline: a comparison of African American and White elders. *Journal of Aging and Health*, *9*(2), 171–184. http://dx.doi.org/10.1177/089826439700900202.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. http://dx.doi.org/10.1023/A:1010933404324.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson …, M., & Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. http://dx.doi.org/10.1038/s41562-018-0399-z.

Coughlan, G., Coutrot, A., Khondoker, M., Minihane, A. M., Spiers, H., & Hornberger, M. (2019). Toward personalized cognitive diagnostics of at-genetic-risk Alzheimer's disease. *Proceedings of the National Academy of Sciences*, *116*(19), 9285–9292. http://dx.doi.org/10.1073/pnas.1901600116.

Coughlan, G., Laczó, J., Hort, J., Minihane, A. M., & Hornberger, M. (2018). Spatial navigation deficits—overlooked cognitive marker for preclinical Alzheimer disease? *Nature Reviews Neurology*, *14*(8), 496–506. http://dx.doi.org/10.1038/s41582-018-0031.

Coutrot, A., Manley, E., Goodroe, S., Gahnstrom, C., Filomena, G., Yesiltepe ., D., & Spiers, H. J. (2022). Entropy of city street networks linked to future spatial navigation ability. *Nature*, *604*(7904), 104–110. http://dx.doi.org/10.1038/s41586-022-04486-7.

Coutrot, A., Schmidt, S., Coutrot, L., Pittman, J., Hong, L., Wiener ., J. M., & Spiers, H. J. (2019). Virtual navigation tested on a mobile app is predictive of real-world wayfinding navigation performance. *PLoS One*, *14*(3), Article e0213272. http://dx.doi.org/10.1371/journal.pone.0213272.

Coutrot, A., Silva, R., Manley, E., de Cothi, W., Sami, S., Bohbot ., V. D., & Spiers, H. J. (2018). Global determinants of navigation ability. *Current Biology*, *28*(17), 2861–2866. http://dx.doi.org/10.1016/j.cub.2018.06.009.

Donald Heth, C., Cornell, E. H., & Flood, T. L. (2002). Self-ratings of sense of direction and route reversal performance. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *16*(3), 309–324. http://dx.doi.org/10.1002/acp.795.

Epstein, R. A., Higgins, J. S., & Thompson-Schill, S. L. (2005). Learning places from views: variation in scene processing as a function of experience and navigational ability. *Journal of Cognitive Neuroscience*, *17*(1), 73–83. http://dx.doi.org/10.1162/0898929052879987.

Farah, M. J. (2017). The neuroscience of socioeconomic status: correlates, causes, and consequences. *Neuron*, *96*(1), 56–71. http://dx.doi.org/10.1016/j.neuron.2017.08.034.

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, *21*, 768–769. http://dx.doi.org/10.1109/tit.1982.1056489.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*(1), 3–42. http://dx.doi.org/10.1007/s10994-006-6226-1.

Grandner, M. A., & Drummond, S. P. (2007). Who are the long sleepers? Towards an understanding of the mortality relationship. *Sleep Medicine Reviews*, *11*(5), 341–360. http://dx.doi.org/10.1016/j.smrv.2007.03.010.

He, C., & Hegarty, M. (2020). How anxiety and growth mindset are linked to navigation ability: Impacts of exploration and GPS use. *Journal of Environmental Psychology*, *71*, Article 101475. http://dx.doi.org/10.1016/j.jenvp.2020.101475.

Hegarty, M., Burte, H., & Boone, A. P. (2018). Individual differences in large-scale spatial abilities and strategies. In D. Montello (Ed.), *Behavioral and cognitive geography* (pp. 231–246). Edward Elgar Publishing, http://dx.doi.org/10.4337/9781784717544.

Hegarty, M., He, C., Boone, A. P., Yu, S., Jacobs, E. G., & Chrastil, E. R. (2022). Understanding differences in wayfinding strategies. *Topics in Cognitive Science*, http://dx.doi.org/10.1111/tops.12592.

Hegarty, M., Montello, D. R., Richardson, A. E., Ishikawa, T., & Lovelace, K. (2006). Spatial abilities at different scales: Individual differences in aptitude-test performance and spatial-layout learning. *Intelligence*, *34*(2), 151–176. http://dx.doi.org/10.1016/j.intell.2005.09.005.

Hegarty, M., Richardson, A. E., Montello, D. R., Lovelace, K., & Subbiah, I. (2002). Development of a self-report measure of environmental spatial ability. *Intelligence*, *30*(5), 425–447. http://dx.doi.org/10.1016/S0160-2896(02)00116-2.

Hejtmánek, L., Oravcová, I., Motýl, J., Horáček, J., & Fajnerová, I. (2018). Spatial knowledge impairment after GPS guided navigation: Eye-tracking study in a virtual town. *International Journal of Human-Computer Studies*, *116*, 15–24. http://dx.doi.org/10.1016/j.ijhcs.2018.04.006.

Hund, A. M., & Padgitt, A. J. (2010). Direction giving and following in the service of wayfinding in a complex indoor environment. *Journal of Environmental Psychology*, *30*(4), 553–564. http://dx.doi.org/10.1016/j.jenvp.2010.01.002.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), Article e124. http://dx.doi.org/10.1371/journal.pmed.1004085.

Janitza, S., Tutz, G., & Boulesteix, A. L. (2016). Random forest for ordinal responses: prediction and variable selection. *Computational Statistics & Data Analysis*, *96*, 57–73. http://dx.doi.org/10.1016/j.csda.2015.10.005.

Jonker, C., Launer, L. J., Hooijer, C., & Lindeboom, J. (1996). Memory complaints and memory impairment in older individuals. *Journal of the American Geriatrics Society*, *44*(1), 44–49. http://dx.doi.org/10.1111/j.1532-5415.1996.tb05636.x.

Jupp, V. (2006). Self-report study. In *The SAGE dictionary of social research methods*. SAGE Publications, Ltd, http://dx.doi.org/10.4135/9780857020116.N186.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, *39*(1), 31–36. http://dx.doi.org/10.1007/bf02291575.

Kansal, T., Bahuguna, S., Singh, V., & Choudhury, T. (2018). Customer segmentation using K-means clustering. In *2018 International conference on computational techniques, electronics and mechanical systems* (pp. 135–139). IEEE, http://dx.doi.org/10.1109/ctems.2018.8769171.

Kochenderfer, M. J., & Wheeler, T. A. (2019). *Algorithms for optimization*. Mit Press.

Kunz, L., Schröder, T. N., Lee, H., Montag, C., Lachmann, B., Sariyska ., R., & Axmacher, N. (2015). Reduced grid-cell–like representations in adults at genetic risk for Alzheimer's disease. *Science*, *350*(6259), 430–433. http://dx.doi.org/10.1126/science.aac8128.

Lester, A. W., Moffat, S. D., Wiener, J. M., Barnes, C. A., & Wolbers, T. (2017). The aging navigational system. *Neuron*, *95*(5), 1019–1035. http://dx.doi.org/10.1016/j.neuron.2017.06.037.

Li, R., & Klippel, A. (2016). Wayfinding behaviors in complex buildings: The impact of environmental legibility and familiarity. *Environment and Behavior*, *48*(3), 482–510. http://dx.doi.org/10.1177/0013916514550243.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129–137. http://dx.doi.org/10.1109/tit.1982.1056489.

McKinlay, R. (2016). Technology: Use or lose our navigation skills. *Nature*, *531*(7596), 573–575. http://dx.doi.org/10.1038/531573a.

Meneghetti, C., Borella, E., Pastore, M., & De Beni, R. (2014). The role of spatial abilities and self-assessments in cardinal point orientation across the lifespan. *Learning and Individual Differences*, *35*, 113–121. http://dx.doi.org/10.1016/j.lindif.2014.07.006.

Montello, D. R. (2005). Navigation. In P. Shah, & A. Miyake (Eds.), *The cambridge handbook of visuospatial thinking* (pp. 257–294). Cambridge, UK: Cambridge University Press, http://dx.doi.org/10.1017/CBO9780511610448.

Namvar, M., Gholamian, M. R., & KhakAbi, S. (2010). A two phase clustering method for intelligent customer segmentation. In *2010 International conference on intelligent systems, modelling and simulation* (pp. 215–219). IEEE, http://dx.doi.org/10.1109/isms.2010.48.

Nazareth, A., Huang, X., Voyer, D., & Newcombe, N. (2019). A meta-analysis of sex differences in human navigation skills. *Psychonomic Bulletin & Review*, *26*(5), 1503–1528. http://dx.doi.org/10.3758/s13423-019-01633-6.

Pagkratidou, M., Galati, A., & Avraamides, M. (2020). Do environmental characteristics predict spatial memory about unfamiliar environments? *Spatial Cognition & Computation*, *20*(1), 1–32. http://dx.doi.org/10.1080/13875868.2019.1676248.

Patel, S. R., Malhotra, A., White, D. P., Gottlieb, D. J., & Hu, F. B. (2006). Association between reduced sleep and weight gain in women. *American Journal of Epidemiology*, *164*(10), 947–954. http://dx.doi.org/10.1093/aje/kwj280.

Pazzaglia, F., Meneghetti, C., Labate, E., & Ronconi, L. (2016). Are wayfinding self-efficacy and pleasure in exploring related to shortcut finding? A study in a virtual environment. In *Spatial cognition X* (pp. 55–68). Cham: Springer, http://dx.doi.org/10.1007/978-3-319-68189-4_4.

Poeppl, T. B., Dimas, E., Sakreida, K., Kernbach, J. M., Markello, R. D., Schöffski ., O., & Bzdok, D. (2022). Pattern learning reveals brain asymmetry to be linked to socioeconomic status. *Cerebral Cortex Communications*, http://dx.doi.org/10.1093/texcom/tgac020.

Puthusseryppady, V., Morrissey, S., Spiers, H., Patel, M., & Hornberger, M. (2022). Predicting real world spatial disorientation in Alzheimer's disease patients using virtual reality navigation tests. *Scientific Reports*, *12*(1), 1–12. http://dx.doi.org/10.1038/s41598-022-17634-w.

Reychav, I., Beeri, R., Balapour, A., Raban, D. R., Sabherwal, R., & Azuri, J. (2019). How reliable are self-assessments using mobile technology in healthcare? The effects of technology identity and self-efficacy. *Computers in Human Behavior*, *91*, 52–61. http://dx.doi.org/10.1016/j.chb.2018.09.024.

Schooler, J. (2011). Unpublished results hide the decline effect. *Nature*, *470*(7335), 437. http://dx.doi.org/10.1038/470437a.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. http://dx.doi.org/10.1177/0956797611417632.

Slavin, M. J., Brodaty, H., Kochan, N. A., Crawford, J. D., Trollor, J. N., Draper, B., & Sachdev, P. S. (2010). Prevalence and predictors of subjective cognitive complaints in the Sydney Memory and Ageing Study. *The American Journal of Geriatric Psychiatry*, *18*(8), 701–710. http://dx.doi.org/10.1097/JGP.0b013e3181df49fb.

Spiers, H. J., Coutrot, A., & Hornberger, M. (2021). Explaining world-wide variation in navigation ability from millions of people: Citizen science project sea hero quest. *Topics in Cognitive Science*, http://dx.doi.org/10.1111/tops.12590.

Taylor, J. L., Miller, T. P., & Tinklenberg, J. R. (1992). Correlates of memory decline: a 4-year longitudinal study of older adults with memory complaints. *Psychology and Aging*, *7*(2), 185. http://dx.doi.org/10.1037/0882-7974.7.2.185.

van der Ham, I. J., van der Kuil, M. N., & Claessen, M. H. (2021). Quality of self-reported cognition: Effects of age and gender on spatial navigation self-reports. *Aging & Mental Health*, *25*(5), 873–878. http://dx.doi.org/10.1080/13607863.2020.1742658.

Van't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology*, *67*, 2–12. http://dx.doi.org/10.1016/j.jesp.2016.03.004.

Wasef, S., Laksono, I., Kapoor, P., Tang-Wei, D., Gold, D., Saripella ., A., & Chung, F. (2021). Screening for subjective cognitive decline in the elderly via subjective cognitive complaints and informant-reported questionnaires: a systematic review. *BMC Anesthesiology*, *21*(1), 1–9. http://dx.doi.org/10.1186/s12871-021-01493-5.

Weisberg, S. M., & Newcombe, N. S. (2018). Cognitive maps: Some people make them, some people struggle. *Current Directions in Psychological Science*, *27*(4), 220–226. http://dx.doi.org/10.1177/0963721417744521.

Wolbers, T., & Hegarty, M. (2010). What determines our navigational abilities? *Trends in Cognitive Sciences*, *14*(3), 138–146. http://dx.doi.org/10.1016/j.tics.2010.01.001.

Wu, S., Yau, W. C., Ong, T. S., & Chong, S. C. (2021). Integrated churn prediction and customer segmentation framework for telco business. *IEEE Access*, *9*, 62118–62136. http://dx.doi.org/10.1109/access.2021.3073776.

Yu, S., Boone, A. P., He, C., Davis, R. C., Hegarty, M., Chrastil, E. R., & Jacobs, E. G. (2021). Age-related changes in spatial navigation are evident by midlife and differ by sex. *Psychological Science*, *32*(5), 692–704. http://dx.doi.org/10.31234/osf.io/rsfpz.