Understanding Preference for Solitude: A Data-Driven Approach Based on a Dual-Process Architecture

TzeHoung Lee, Singapore University of Social Sciences, Singapore
Peter Tay, Singapore Institute of Technology, Singapore

The Asian Conference on Aging & Gerontology 2026
Official Conference Proceedings

Abstract

Why do some people actively seek time alone while others prefer constant social contact? We examined this question using behavioral data from 203 older adults in Singapore, measuring 104 aspects of their lives—from daily activity patterns to personality traits to health status. We compared three analytical approaches: (1) traditional factor analysis with ordinary least squares regression, (2) confirmatory factor analysis for construct validation, and (3) a two-stage ensemble method inspired by dual-process theories of cognition. The two-stage approach explained 28% of variance in solitude preference, substantially outperforming traditional methods (13%) and single-model machine learning (21%). Critically, proper hyperparameter tuning revealed that the analytical refinement stage (gradient boosting) added meaningful value beyond initial pattern recognition (random forest), increasing explained variance by 7 percentage points. The strongest predictors were behavioral patterns (hours spent alone, solitary activities) rather than personality traits—extraversion ranked only 15th among 104 predictors. Confirmatory factor analysis validated solitude preference as a distinct construct, separating from loneliness ($r = .42$), social anxiety ($r = .18$), and extraversion ($r = -.36$). These findings advance understanding of solitude as a meaningful individual difference with practical implications for well-being in later life and demonstrate methodological advantages of ensemble methods for complex behavioral data with mixed variable types.

Keywords: solitude preference, dual-process theory, individual differences, factor analysis, ensemble methods, older adults

Introduction

Background and Motivation

Humans are fundamentally social creatures. We form relationships, seek companionship, and generally thrive on social connection. Yet some people actively choose to spend time alone—not because they lack social skills or suffer from loneliness, but because they genuinely enjoy solitude and find it restorative. Understanding why some people prefer solitude while others seek constant social interaction matters for several reasons.

Americans spend substantial time alone—17% for children to 48% for retirees (Larson, 1990)—which can be enriching or isolating depending on individual preference. Society often stigmatizes solitude, equating it with loneliness, yet understanding healthy solitude is increasingly important for aging populations.

The key insight from previous research: not everyone who spends time alone is lonely, and not everyone who prefers solitude is avoiding others (Burger, 1995). Some people have learned to appreciate the benefits that solitude brings—time for self-reflection, creative pursuits, emotional renewal, and personal restoration. Distinguishing between healthy solitude preference and problematic social isolation requires understanding what predicts this individual difference.

What We Know About Solitude

Previous research shows that solitude can have both positive and negative effects. On the negative side, excessive isolation associates with loneliness, depression, and poorer physical health. People who spend most of their time alone often report feeling bored, passive, and emotionally flat. Adolescents who withdraw socially may struggle with identity development and peer relationships.

On the positive side, self-imposed solitude can promote well-being. Periods alone allow people to engage in necessary self-reflection, work through personal problems, explore creative interests, and restore emotional energy after demanding social interactions. Historical figures from Carl Jung to Virginia Woolf deliberately sought extended periods of isolation to deepen their intellectual and creative work. Abraham Maslow (1970) found that psychologically healthy, self-actualized people often expressed strong preferences for privacy and solitude while also maintaining deep friendships and warm interpersonal relationships. Their preference for solitude reflected self-knowledge and personal growth, not social anxiety or interpersonal problems.

Research Challenges

Understanding who prefers solitude requires addressing conceptual and methodological challenges.

Methodologically, behavioral datasets combine numerical and categorical information, and traditional methods struggle with this combination.

Moreover, traditional regression approaches assume linear, additive relationships: each predictor contributes independently to the outcome. But behavioral phenomena often involve interactions and nonlinearities. The effect of age on solitude preference might differ for men versus women, or health status might matter more for people living alone. Traditional methods require researchers to specify these interactions in advance; data-driven methods can discover them automatically.

Study Objectives

This study addresses three objectives. First, we validate preference for solitude as a distinct construct using confirmatory factor analysis, testing whether it separates empirically from loneliness, social anxiety, and extraversion. Second, we compare traditional analytical methods (factor analysis with ordinary least squares regression) against modern ensemble approaches for predicting individual differences in solitude preference. Third, we test whether a dual-process analytical architecture—combining parallel pattern recognition with sequential analytical refinement—provides advantages for complex behavioral data with mixed variable types and potential nonlinear relationships.

Theoretical Framework: Dual-Process Architecture

We drew inspiration from dual-process theories in cognitive psychology (Evans & Stanovich, 2013; Kahneman, 2011). These theories propose that human thinking operates through two complementary systems.

System 1 (Fast Thinking) provides rapid, automatic, intuitive processing by quickly recognizing familiar patterns but can make systematic errors.

System 2 (Slow Thinking) provides deliberate, analytical, controlled processing, working step-by-step to check and correct System 1's impressions.

Effective thinking combines both systems.

We designed a two-stage analytical method mirroring this architecture: Stage 1 (Random Forest) captures broad patterns through parallel processing across 300 trees; Stage 2 (Gradient Boosting) corrects systematic biases through sequential processing.

Critical question: Does Stage 2 actually add value when properly configured, or does Stage 1 capture all relevant patterns?

Methods

Participants

We analyzed data from 203 community-dwelling older adults in Singapore (*M* age = 68 years, range 55–89; 58% female). All participants were healthy, independent individuals living in the community (not in institutional care). They completed comprehensive assessments of their psychological characteristics, social relationships, health, and daily activities. The study received ethical approval and all participants provided informed consent.

Measures

Preference for Solitude: We used Burger's (1995) Preference for Solitude Scale, a well-validated 12-item forced-choice measure. Each item presents two statements, and participants choose which better describes them. Example items: "I try to structure my day so that I always have some time to myself" versus "I try to structure my day so that I always am doing something with someone"; "I like to vacation in places where there are few people around and a lot of serenity and quiet" versus "I like to vacation in places where there are a lot of people around and a lot of activities going on"; "After spending a few hours surrounded by a lot of people, I am usually eager to get away by myself" versus "After spending a few hours surrounded by a lot of people, I usually find myself energized and stimulated."

Scores range from 0 (strong preference for social contact) to 12 (strong preference for solitude). Internal consistency in our sample: $\alpha = .71$. Sample distribution: $M = 4.9$, $SD = 2.6$, meaning slightly more people preferred some solitude over constant social contact, with substantial individual variation.

Importantly, this scale measures preference—how much people want to spend time alone when given the choice. It doesn't measure loneliness (feeling isolated and wanting more social contact) or social anxiety (avoiding people due to fear). Previous research shows that people can score high on preference for solitude while also having close friendships and low social anxiety.

Predictor Variables (104 total): We examined 104 different aspects of people's lives, grouped into five categories.

Demographics and Social Context (15 variables): Age, gender (male/female), ethnicity (Chinese/Malay/Indian/Other), marital status (single/married/divorced/widowed/separated), education level (primary/secondary/junior college/polytechnic/university/postgraduate), current living arrangement (alone/with spouse/with children/with others), residential type (public housing/condominium/landed property), whether they employ a foreign domestic worker, height, weight, and household income.

Social Network Characteristics (6 variables): Number of close friends, frequency of family contact, participation in organized social activities (clubs, religious groups, volunteer organizations), social support received from others, social support provided to others, and diversity of their social network (how many different types of people they interact with regularly).

Daily Activity Patterns (22 variables): Hours per week spent alone, with friends, with family, in organized group activities, in solitary leisure (reading, watching television, hobbies), in social leisure (dining out, visiting friends, attending events), and in physical activity. We measured these separately for weekdays versus weekends, yielding 22 specific activity variables.

Health and Functioning (26 variables): Nineteen specific chronic health conditions (hypertension, diabetes, cardiovascular disease, stroke history, arthritis, osteoporosis, respiratory conditions, cancer history, kidney disease, liver disease, and others) plus seven health indicators (self-rated health, physical functioning, activities of daily living, instrumental activities of daily living, pain severity, sleep quality, number of medications taken).

Psychological Characteristics (35 variables): Big Five personality traits (extraversion, neuroticism, openness to experience, agreeableness, conscientiousness) measured using the NEO Five-Factor Inventory; loneliness (UCLA Loneliness Scale, 3-item version); depression symptoms (CES-D short form); life satisfaction; sense of meaning in life; resilience; social anxiety measured via the Interaction Anxiousness Scale; and various other validated psychological constructs.

Analytical Strategy

We compared three analytical approaches to understand which best predicts individual differences in solitude preference.

Approach 1: Traditional Factor Analysis With OLS Regression

This represents the standard approach in behavioral science when faced with many predictors.

Step 1 – Dimensionality Reduction: We applied principal components analysis to the 104 predictors. Numerical variables were standardized (mean = 0, SD = 1). Categorical variables were converted using one-hot encoding (creating binary dummy variables for each category). We retained components with eigenvalues greater than 1.0, yielding 23 principal components that collectively explained 71% of total variance in the predictor space.

Step 2 – Ordinary Least Squares Regression: We regressed solitude preference scores onto the 23 principal components using standard OLS regression. This produces linear predictions where each component contributes additively to the outcome.

Approach 2: Construct Validation via Exploratory and Confirmatory Factor Analysis

To test whether preference for solitude represents a distinct construct or simply reflects introversion, loneliness, or social anxiety, we conducted both exploratory and confirmatory factor analyses.

Exploratory Factor Analysis: We conducted EFA on items from four related scales: Burger's Preference for Solitude Scale (12 items), UCLA Loneliness Scale (3 items), Social Anxiety subscale from the Interaction Anxiousness Scale (5 items), and Extraversion subscale from the NEO Five-Factor Inventory (12 items). This yielded 32 items total. We used maximum likelihood extraction with oblique rotation (promax) to allow factors to correlate naturally. We tested solutions from 1 factor through 4 factors, evaluating each using eigenvalues, scree plots, and interpretability.

Confirmatory Factor Analysis: Based on EFA results and theoretical expectations, we tested a 4-factor model explicitly specifying: (1) Preference for Solitude factor (12 items from Burger scale), (2) Loneliness factor (3 items), (3) Social Anxiety factor (5 items), (4) Extraversion factor (12 items). We evaluated model fit using multiple indices: Comparative Fit Index (CFI $>$ .90 indicates acceptable fit), Tucker-Lewis Index (TLI $>$ .90), Root Mean Square Error of Approximation (RMSEA $<$ .08), and Standardized Root Mean Square Residual (SRMR $<$ .08). We also tested alternative models (1-factor, 3-factor) to verify that the 4-factor structure fit better than simpler alternatives.

Approach 3: Two-Stage Ensemble Method

Stage 1 – Random Forest (System 1 Analogue): We trained a random forest with the following specifications: 300 trees, max features per split = $\sqrt{p}$ (square root of number of predictors, ensuring diversity across trees), min samples per split = 10 (preventing overly specific splits), min samples per leaf = 5 (ensuring each terminal node represents generalizable patterns), bootstrap sampling enabled (each tree sees different data sample), out-of-bag scoring enabled (internal validation).

Categorical variables (gender, ethnicity, marital status, education level, living arrangement, residential type) were handled natively through binary splits without requiring numerical encoding. This eliminates researcher decisions about how to encode categories.

Stage 2 – Gradient Boosting (System 2 Analogue): We applied XGBoost to the residual errors from Stage 1—the differences between actual solitude preference scores and Stage 1 predictions. Initial implementations used default hyperparameters and showed minimal improvement (adding only 0.2 percentage points of explained variance), suggesting the approach was not working as intended.

We conducted systematic hyperparameter tuning using 5-fold cross-validation to optimize Stage 2 performance. We tested the following parameter ranges: learning rate [0.001, 0.01, 0.05, 0.1], max depth [3, 4, 5, 6], number of estimators [50, 100, 150, 200], min child weight [1, 3, 5], subsample [0.7, 0.8, 0.9, 1.0], colsample by tree [0.7, 0.8, 0.9, 1.0].

Optimal configuration identified through cross-validation: learning rate = 0.01 (slow, gradual corrections), max depth = 4 (shallow trees focusing on systematic patterns), number of estimators = 150 (sufficient iterations to learn corrections), min child weight = 3 (regularization preventing overfitting), subsample = 0.8 (using 80% of data per tree), colsample by tree = 0.8 (using 80% of features per tree), early stopping = 10 rounds (halt training when validation performance stops improving).

With optimized hyperparameters, Stage 2 substantially improved predictions, increasing explained variance from 21% (Stage 1 only) to 28% (both stages combined)—a 7 percentage point improvement that validates the dual-process architecture when properly configured.

Validation Strategy

We used 10-fold cross-validation with stratified sampling (stratifying by solitude preference quartiles to ensure balanced representation of the full range). For each fold: train on 90% of data, predict on held-out 10%, calculate R^2 (proportion of variance explained) and RMSE (root mean squared error in scale points).

We repeated this entire process 10 times with different random seeds, yielding 100 total train-test evaluations (10 folds × 10 random seeds). This provides robust estimates of generalization performance. We report mean performance metrics and 95% confidence intervals computed via bootstrap.

Supplementary Analysis: Addressing Circularity Concerns

Behavioral time-use variables (hours spent alone, hours in solitary leisure activities) partially overlap with the outcome we're predicting (preference for solitude)—they represent behavioral manifestations of the very preference we're trying to predict. To address concerns that our findings might be circular, we conducted supplementary analyses excluding these variables.

We removed 8 directly behavioral predictors: hours spent alone on weekdays, hours spent alone on weekends, hours in solitary leisure activities, hours with friends on weekdays, hours with friends on weekends, hours with family on weekdays, hours with family on weekends, and hours in organized group activities. This left 96 predictors assessing person characteristics and life circumstances (demographics, social network structure, health status, psychological characteristics) rather than direct behavioral manifestations of solitude preference.

Results

Construct Validation: Exploratory and Confirmatory Factor Analysis

Exploratory Factor Analysis: Parallel analysis and scree plot inspection both suggested a 4-factor solution as optimal. The 4-factor model with oblique rotation showed clean factor structure with minimal cross-loadings.

Factor 1 (Preference for Solitude): All 12 items from Burger's Preference for Solitude Scale loaded primarily on this factor, with loadings ranging from .45 to .72. Eigenvalue = 6.8, explaining 21% of total variance.

Factor 2 (Extraversion): All 12 items from the NEO Extraversion subscale loaded on this factor, with loadings from .52 to .81. Eigenvalue = 5.3, explaining 16% of variance.

Factor 3 (Social Anxiety): All 5 items from the Interaction Anxiousness Scale loaded on this factor, with loadings from .61 to .78. Eigenvalue = 3.2, explaining 10% of variance.

Factor 4 (Loneliness): All 3 UCLA Loneliness Scale items loaded on this factor, with loadings from .72 to .85. Eigenvalue = 2.4, explaining 7% of variance.

Cross-loadings were minimal (all below .30), indicating good discriminant validity. Items loaded cleanly on their theoretically expected factors.

Interfactor correlations revealed meaningful but moderate relationships: Preference for Solitude ↔ Extraversion ($r = -.36$, indicating introverts somewhat more likely to prefer solitude but relationship far from perfect); Preference for Solitude ↔ Social Anxiety ($r = .18$, weak positive relationship); Preference for Solitude ↔ Loneliness ($r = .42$, moderate positive relationship but not redundant); Extraversion ↔ Social Anxiety ($r = -.52$); Extraversion ↔ Loneliness ($r = -.44$); Social Anxiety ↔ Loneliness ($r = .38$).

These correlations confirm that preference for solitude, while related to extraversion and loneliness, represents a distinct construct rather than simply the inverse of sociability or a manifestation of loneliness.

Confirmatory Factor Analysis: The hypothesized 4-factor model showed acceptable to good fit across multiple indices: $\chi^2(458) = 687.3, p < .001$; CFI = .92 (above .90 threshold); TLI = .91 (above .90 threshold); RMSEA = .051 with 90% confidence interval [.044, .058] (well below .08 threshold); SRMR = .062 (below .08 threshold).

All factor loadings were statistically significant ($p < .001$) and in the expected directions. The pattern of loadings matched theoretical expectations and EFA results.

We tested alternative models to verify that the 4-factor structure provided superior fit compared to simpler alternatives. A 1-factor model collapsing all items onto a general "social preference" factor showed poor fit: CFI = .61, RMSEA = .128 (substantially worse). A 3-factor model combining preference for solitude with extraversion showed poor fit: CFI = .78, RMSEA = .095. A 3-factor model combining preference for solitude with loneliness also showed poor fit: CFI = .81, RMSEA = .088.

The 4-factor model showed significantly better fit than all alternatives (chi-square difference tests: all $p < .001$), providing strong evidence that preference for solitude functions as a distinct construct worthy of independent study.

Comparative Predictive Performance Across Methods

Table 1
Predictive Performance Across Analytical Methods

Method	R^2	RMSE	95% CI for R^2
Traditional PCA + OLS Regression	.13	2.42	[.09, .17]
Random Forest (Stage 1 only)	.21	2.31	[.17, .25]
Two-Stage (initial hyperparameters)	.21	2.30	[.17, .25]
Two-Stage (optimized hyperparameters)	.28	2.20	[.24, .32]

Note. R^2 = proportion of variance explained in solitude preference scores. RMSE = root mean squared error in scale points (0–12 range). Results averaged across 10-fold cross-validation with 10 random seeds (100 total train-test evaluations). 95% confidence intervals computed via bootstrap resampling.

Three key findings emerged from this comparison:

Finding 1: The two-stage ensemble method with optimized hyperparameters substantially outperformed traditional factor analysis plus regression (.28 versus .13, representing a 115% improvement in explained variance). In concrete terms: if you know nothing about someone, your best guess for their solitude preference would be the sample average (around 5 on the 0–12 scale), and you'd typically be off by about 2.6 points. Traditional methods reduce this error to 2.42 points. Our optimized two-stage approach reduces it further to 2.20 points—a meaningful practical improvement.

Finding 2: Proper hyperparameter tuning proved critical for the dual-process architecture. Initial implementations using default hyperparameters showed Stage 2 adding minimal value (increasing R^2 from .21 to only .21, essentially no improvement). After systematic tuning via cross-validation, Stage 2 provided substantial improvement (increasing R^2 from .21 to .28, a 7 percentage point gain). This demonstrates that simply "using machine learning" is insufficient—careful configuration determines whether methods achieve their potential.

Finding 3: Even Random Forest alone (.21) outperformed traditional methods (.13), likely due to three factors: (a) native handling of categorical variables without arbitrary numerical encoding decisions, (b) automatic discovery of nonlinear relationships and interactions without requiring researcher specification, (c) robustness through bootstrap aggregation across 300 diverse decision trees.

Why Stage 2 Initially Failed and How We Systematically Fixed It

Initial gradient boosting implementations used default hyperparameters optimized for large datasets ($N > 10{,}000$). Our dataset ($N = 203$) required different configuration. We identified four specific problems and their solutions through systematic investigation:

Solution: We reduced the learning rate to 0.01, forcing more gradual error correction. With slower learning, the algorithm identifies robust patterns that replicate across data samples rather than memorizing training-specific quirks.

Solution: We reduced max depth to 4, forcing simpler decision rules. Shallow trees capture broad systematic patterns while avoiding overfitting to training-specific details.

Solution: We increased min child weight to 3, requiring that each split represent patterns across multiple people rather than individual-specific quirks. This regularization improves generalization.

Solution: We implemented early stopping with a patience of 10 rounds—if validation error doesn't improve for 10 consecutive iterations, training halts. This prevents overfitting while allowing sufficient iterations to learn genuine corrections.

Residual Analysis Reveals What Stage 2 Corrects: To understand what Stage 2 actually learns, we analyzed systematic patterns in Stage 1's prediction errors. We found that Stage 1 made consistent mistakes for certain subgroups:

Underestimation Pattern: Stage 1 consistently underestimated solitude preference for highly educated individuals (postgraduate degree) living alone. For this subgroup, mean residual error = +1.2 scale points, indicating Stage 1 predictions were systematically too low.

Overestimation Pattern: Stage 1 consistently overestimated solitude preference for married individuals with large social networks (6+ close friends). For this subgroup, mean residual error = −.9 scale points, indicating Stage 1 predictions were systematically too high.

Stage 2's role: With optimized hyperparameters, gradient boosting successfully identified these systematic biases and built correction rules. Analysis of Stage 2 trees revealed splits like "if education = postgraduate AND living arrangement = alone, then add correction of +0.7" and "if marital status = married AND number of close friends > 6, then add correction of −.5." Stage 2 corrected approximately 60% of the systematic bias, substantially improving predictions for these subgroups.

This demonstrates that the dual-process architecture provides genuine value when properly configured: broad pattern recognition (Stage 1) captures obvious relationships efficiently through parallel processing, while focused analytical refinement (Stage 2) corrects subtle systematic errors through sequential processing.

Feature Importance Analysis: What Predicts Solitude Preference?

Table 2
Top 20 Predictors of Preference for Solitude

Rank	Predictor	Type	Importance
1	Hours spent alone (weekdays)	Behavioral	.082
2	Hours spent alone (weekends)	Behavioral	.071
3	Time in solitary leisure (reading)	Behavioral	.058
4	Hours with friends (weekdays)	Behavioral	.052
5	Age	Demographic	.048
6	Participation in organized activities	Behavioral	.041
7	Diversity of social network	Social	.037
8	Hours in group activities	Behavioral	.035
9	Self-rated health	Health	.033
10	Number of close friends	Social	.031
11	Life satisfaction	Psychological	.029
12	Hours with family (weekends)	Behavioral	.027
13	Living arrangement	Demographic	.026
14	Education level	Demographic	.024
15	Extraversion	Psychological	.023
16	Hours watching television	Behavioral	.022
17	Social support received	Social	.021
18	Hours dining out	Behavioral	.019
19	Family contact frequency	Social	.018
20	Loneliness	Psychological	.017

Note. Importance scores derived from permutation importance methodology averaged across both ensemble stages. Behavioral activity patterns comprise 11 of the top 20 predictors. Extraversion—theoretically expected to be the primary predictor—ranked only 15th among 104 variables.

Several striking patterns emerge from the feature importance analysis. First, actual behavioral patterns (hours spent alone, time in solitary activities, participation in social versus solitary pursuits) emerged as the strongest predictors of solitude preference. What people actually do with their time reveals their preferences more clearly than self-reported personality traits or other individual characteristics.

Second, extraversion ranked only 15th among our 104 predictors despite being theoretically central to solitude preference. In traditional personality psychology, we would expect extraversion to be the primary predictor. This finding has two possible interpretations: (1) Behavioral redundancy—once you know how much time someone actually spends alone and in solitary activities, knowing whether they self-identify as extraverted or introverted adds little additional predictive information; or (2) Conceptual distinctness—preference for solitude may genuinely differ from the introversion-extraversion dimension.

Third, life circumstances matter. Age (ranking 5th), self-rated health (9th), living arrangement (13th), and education level (14th) all emerged as important predictors. This suggests that solitude preference reflects not just stable personality traits but also life stage, physical capabilities, and social circumstances.

Fourth, no single factor dominated. Instead, solitude preference emerged from combinations of demographics, health status, social context, activity patterns, and psychological characteristics, suggesting multiple pathways to preferring solitude.

Supplementary Analysis: Excluding Behavioral Time-Use Variables

Table 3
Performance When Excluding Behavioral Time-Use Variables

Method	Full Model (104)	Reduced Model (96)
Traditional PCA + OLS	.13	.11
Random Forest only	.21	.16
Two-Stage (optimized)	.28	.19

Note. Full model includes all 104 predictors. Reduced model excludes 8 behavioral time-use variables (hours alone on weekdays/weekends, hours with friends on weekdays/weekends, hours with family on weekdays/weekends, hours in solitary leisure, hours in organized group activities).

Three findings emerged from this supplementary analysis:

First, excluding behavioral variables reduced performance across all methods, confirming that time-use patterns strongly predict solitude preference. This makes conceptual sense—people who prefer solitude are more likely to structure their lives to include time alone.

Second, even without behavioral variables, the optimized two-stage approach (.19) substantially outperformed traditional methods (.11), demonstrating that the ensemble method's advantages extend beyond simply including time-use data. The method still benefits from native categorical handling, automatic interaction detection, and robustness through aggregation.

Third, in the reduced model, personality traits gained importance. Extraversion moved from 15th to 3rd in importance rankings, neuroticism moved from 24th to 7th, and loneliness moved from 20th to 5th. This demonstrates that our main finding (behavioral patterns outperform personality traits) partially reflects behavioral variables obscuring personality's role. When behavioral manifestations are removed, underlying personality characteristics become more predictive.

This supplementary analysis reveals important nuance: our primary models achieve strong performance partly through near-circular prediction (using behavioral manifestations to predict behavioral preferences), but the ensemble approach provides methodological advantages even when restricted to person characteristics and life circumstances that don't directly reflect solitude behavior.

Automatic Detection of Interactions and Nonlinearities

The ensemble methods automatically discovered several interactions and nonlinearities: Age interacted with living arrangement (age strongly predicted solitude preference among those living alone but not among those living with family); health status interacted with social network size (close friendships negatively predicted solitude preference among healthy individuals but not among those in poor health); education showed a U-shaped pattern (highest solitude preference among those with primary and postgraduate education, lowest among those with secondary education). These patterns would be missed by traditional regression unless specifically hypothesized in advance.

Discussion

This study compared traditional and modern analytical methods for predicting individual differences in preference for solitude among older adults.

Main Finding 1: Preference for Solitude as Distinct Construct

Confirmatory factor analysis validated preference for solitude as a distinct construct, empirically separable from extraversion ($r = -.36$), social anxiety ($r = .18$), and loneliness ($r = .42$). While moderately correlated with these related constructs—as would be expected given conceptual overlap—solitude preference showed sufficient discriminant validity to warrant study as an independent individual difference.

The correlation with extraversion explains only 13% of variance, challenging the assumption that solitude preference simply reflects introversion.

The moderate positive correlation with loneliness indicates overlap but not redundancy. Many people who prefer solitude are not lonely—they have fulfilling social relationships and simply value periods alone. This finding has important theoretical implications. It suggests that preference for solitude involves more than just temperamental introversion. Some extraverts may prefer solitude for restoration after intense social engagement; some introverts may not particularly value time alone despite being quiet and reserved.

Main Finding 2: Substantial Predictive Performance Advantages of Ensemble Methods

The optimized two-stage ensemble approach explained 28% of variance in solitude preference, substantially outperforming traditional factor analysis plus regression (13%) and providing meaningful improvement over single-model machine learning (Random Forest: 21%). This represents a 115% improvement in explained variance over traditional methods and a 33% improvement over the single-model baseline.

Three factors likely contribute to these performance advantages. First, native categorical handling: Traditional approaches required converting gender, ethnicity, marital status, education level, and living arrangement into numbers through one-hot encoding or other schemes. These conversions involve arbitrary decisions that can bias results. Tree-based methods make direct categorical splits without numerical encoding, eliminating researcher degrees of freedom.

Second, automatic discovery of nonlinear relationships and interactions: Our analyses revealed Age × Living Arrangement interactions, Health × Social Network Size interactions, and U-shaped Education effects that traditional regression would miss unless researchers specified them in advance. The ensemble method discovered these patterns directly from data.

Third, robustness through aggregation: Stage 1 averages predictions across 300 diverse decision trees, each trained on different data samples and variable subsets. This ensemble averaging produces predictions more robust to individual tree quirks and more likely to generalize to new people than single-model predictions.

Main Finding 3: The Dual-Process Architecture Requires and Rewards Proper Configuration

Initial implementations showed Stage 2 (gradient boosting) adding minimal value beyond Stage 1 (random forest)—only 0.2 percentage points of explained variance. This initially appeared to contradict our dual-process theoretical framework. However, systematic hyperparameter tuning revealed that Stage 2 could add substantial value (7 percentage points) when properly configured for our dataset characteristics.

The key insight: Stage 2 specializes in correcting systematic biases that Stage 1 makes consistently across cases. This requires specific configuration: slow learning rate (to avoid overfitting to noise in residuals), shallow trees (to focus on systematic patterns rather than individual-specific corrections), regularization (to prevent memorizing training data), and early stopping (to halt when corrections plateau).

Residual analysis revealed what Stage 2 learns: Stage 1 systematically underestimated solitude preference for highly educated individuals living alone (mean error = +1.2 scale points) and overestimated for married individuals with large social networks (mean error = −.9 points). Stage 2 successfully identified these patterns and corrected approximately 60% of systematic bias.

This validates our dual-process architecture when properly tuned: broad pattern recognition (Stage 1) captures obvious relationships efficiently through parallel processing, while focused analytical refinement (Stage 2) corrects subtle systematic errors through sequential processing. The architecture mirrors complementary functions of System 1 and System 2 cognition.

Limitations and Future Directions

Several important limitations qualify our findings and suggest directions for future research.

Sample Size and Generalizability: With N = 203 participants, all from Singapore, all aged 55 and above, we cannot assume findings generalize to younger adults, Western populations, or different cultural contexts. Our effect sizes likely overestimate what would be found in larger, more representative samples. Replication with larger samples across diverse populations is essential.

Cross-Sectional Design Prevents Causal Inference: We measured all variables at a single time point, preventing causal conclusions. Longitudinal designs following people over time could disentangle directional relationships and examine how solitude preferences change across life transitions.

Predictor Overlap with Outcome: Hours spent alone and time in solitary activities partially overlap conceptually with preference for solitude. Our supplementary analysis excluding behavioral variables addresses this partially, but the reduced model explains only 19% of variance.

Substantial Unexplained Variance: Even our best-performing model explains only 28% of variance in solitude preference. This means 72% of individual variation remains unexplained by our 104 measured predictors. Unmeasured factors likely include early life experiences, cultural values, spiritual beliefs, cognitive styles, and inherent randomness in behavioral preferences.

Implications

For Personality Theory: Preference for solitude deserves study as a distinct individual difference that goes beyond standard Big Five personality dimensions.

For Aging Research and Practice: Among older adults in our sample, solitude preference appeared linked to positive characteristics rather than problematic isolation. Aging services should respect individual differences in preferred balance between social contact and solitude.

For Clinical Assessment: Clinicians should carefully distinguish between preference for solitude (potentially healthy) and problematic isolation (loneliness, depression).

For Research Methodology: Our findings demonstrate advantages of ensemble methods but also show that proper hyperparameter tuning determines whether methods achieve their potential.

Conclusion

This study demonstrates that individual differences in preference for solitude can be successfully predicted from comprehensive behavioral data using ensemble methods inspired by dual-process theories of cognition. We report three key contributions.

First, confirmatory factor analysis validated preference for solitude as a distinct psychological construct, empirically separable from extraversion, social anxiety, and loneliness despite moderate correlations. This supports treating solitude preference as a meaningful individual difference worthy of independent theoretical and empirical attention.

Second, a properly configured two-stage ensemble approach substantially outperformed traditional analytical methods (explaining 28% versus 13% of variance), demonstrating practical advantages for complex behavioral datasets containing mixed variable types and nonlinear relationships. The performance gains stem from native categorical handling, automatic interaction detection, and robustness through aggregation.

Third, systematic hyperparameter optimization revealed that analytical refinement (Stage 2 gradient boosting) adds meaningful value beyond pattern recognition (Stage 1 random forest) by correcting systematic prediction biases, validating our dual-process theoretical framework when properly configured. This highlights the critical importance of careful tuning rather than applying default configurations.

These findings advance both theoretical understanding (clarifying solitude preference as conceptually distinct from introversion with multiple behavioral pathways) and methodological practice (demonstrating ensemble advantages while emphasizing proper configuration). Future research should address current limitations through larger and more diverse samples, longitudinal designs that enable causal inference, and careful conceptual distinction between prediction goals and explanation goals.

References

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Burger, J. M. (1995). Individual differences in preference for solitude. *Journal of Research in Personality, 29*(1), 85–108.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 785–794.

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*(3), 223–241.

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time.* Department of Statistics, Columbia University.

Kahneman, D. (2011). *Thinking, fast and slow.* Farrar, Straus and Giroux.

Larson, R. W. (1990). The solitary side of life: An examination of the time people spend alone from childhood to old age. *Developmental Review, 10*(2), 155–183.

Maslow, A. H. (1970). *Motivation and personality* (2nd ed.). Harper & Row.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12*(6), 1100–1122.

Contact email: thlee018@suss.edu.sg