Comparison of Automated Short Form Selection Strategies

Anthony Raborn[1] Walter Leite Katerina Marcoulides

University of Florida

1: Corresponding author: anthony.w.raborn@gmail.com

November 8, 2018

Goals

Goals of this Study

- Compare different scale reduction strategies
  1. Time to converge (faster is better)
  2. Model fit of final scales (better fit is better)
  3. Reliability of final scales (higher reliability is better)
  4. Removal of specific problematic items (fewer problematic items is better)
- Determine which factors effect these comparisons
  1. Population model type (one factor, three factor, bifactor)
  2. Severity of problematic items (none, minor, major)
  3. Strength of relationship to external criterion (none, moderate)

Introduction

Applications of Psychometric Scales

- Applied researchers are often faced with a dilemma:
  - Option A: Use a well-established but lengthy scale with fewer additional items
  - Option B: Use a few items from a scale with more additional items
- Both options have some drawbacks!
  - Option A: Potentially longer administration time for less information
  - Option B: Potentially greater information but weaker validity evidence
- In the literature, researchers attempt to use Option B with some effort spent on buoying the validity evidence

Examples in the Literature

1. Hand-Selecting Items
   - using theoretical or practical justifications per item (e.g., Noble, Jensen, Naylor, Bhullar, & Akeroyd, 2013)
   - Retaining one of many redundant items (e.g., Dennis, 2003)
2. Statistical Criteria
   - Retaining items with high factor loadings or item correlations (e.g., Byrne & Pachana, 2011; Wester, Vogel, O'neil, & Danforth, 2012)
   - Selecting items that improve measures of reliability and/or dimensionality (e.g., Lim & Chapman, 2013; Veale, 2014)

Overall, the focus of the above examples are on the internal structure of the scales. This is in spite of researchers wanting to use the short form for predictive/correlational purposes.

Specific Example: Positive Mental Health Assessment

Petrillo et al.(2015) developed a short form of a positive mental health assessment for Italian respondents. Items were selected from twelve other scales with a focus on the final form's internal structure (a second-order three factor model).

- CFI: .93, TLI: .91, RMSEA: .06[1]
- Overall scale $\alpha = .86$

Despite this, the short form was only weakly correlated with other scales mwith good validity evidence which also assessed mental health.

- Range of total score correlations[2]: 0.20 to 0.62

[1] Adequate, but not ideal, fit (Hu & Bentler, 1999)

[2] Absolute value of correlations. Mean absolute correlation: 0.37

Problem

Creating short forms with (1) good internal structure and (2) good predictive, convergent, and/or divergent validity is difficult by hand.

One potential solution would be to use metaheuristic optimization algorithms (Dréo, Pétrowski, Siarry, & Taillard, 2006). These algorithms can *simultaneously optimize* multiple criteria, particularly the internal structure and external relationships of a scale.

Theoretical Framework

Previous Attempts

There have been a few methods in the literature of using algorithms to shorten scales. Some of the common ones are:

1. "Maximize Main Loadings"
   - An often-used algorithm that essentially automates the process of picking the items with the highest factor loadings. Generally results in a homogenous item structure within each factor (Olaru, Witthöft, & Wilhelm, 2015)
     - Not a metaheuristic algorithm, so not included in study
2. Ant Colony Optimization (ACO)
   - Leite, Huang, & Marcoulides (2008) developed and compared ACO to traditional methods to create a short form of a quality-of-life scale for diabetics while optimizing the relationship with an external criterion variable

Previous Attempts, cont'd

3. Tabu Search (TS)
   - Marcoulides & Drezner (2004) demonstrated a use of TS to reduce the number of items loading on factors
4. Genetic Algorithm (GA)
   - Yarkoni (2010) developed the particular application to combine 203 distinct personality scales into one inventory
     - Did *not* include any external relationships in the algorithm or a way to include them

Ant Colony Optimization Algorithm

ACO (Colorni, Dorigo, Maniezzo, & others, 1992) mimics the behavior of ants searching for the shortest path to a food source.

The ants leave the nest (N) leaving pheremone trails (yellow lines) to the food source (F) that the next iteration of ants follow.

Over time, the pheremone builds up along the shortest path until (almost) all ants follow the same path (see next slide).

Leite et al. (2008) set all paths (items) equal initially, used 20 ants per iteration, used the mean standardized regression coefficients of the model as the pheremone level, and the overall model fit by CFI, TLI, and RMSEA as the food source.

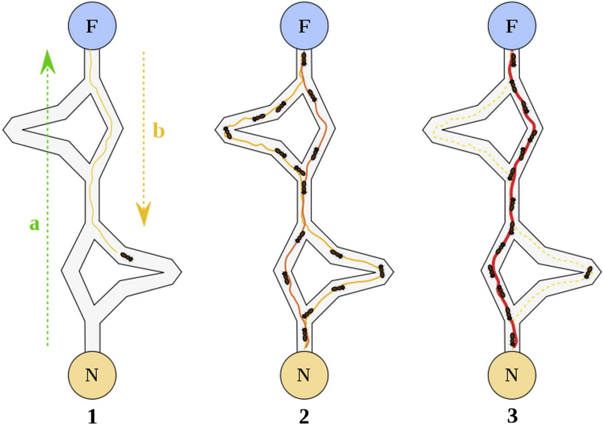

Figure 1: Ant Colony Optimization (Toksari, 2016)

Tabu Search

TS (Glover, 1989) is a local search metaheuristic that can accept potential solutions that are worse than the current solution if no better solutions exist. It employs a list of *tabu* solutions that have already been explored; this list keeps solutions for a certain number of iterations before they are removed from the list.

Within each iteration, TS explores all possible local models (i.e., models that differ by one parameter). Typically, the algorithm stops after a predetermined number of iterations.

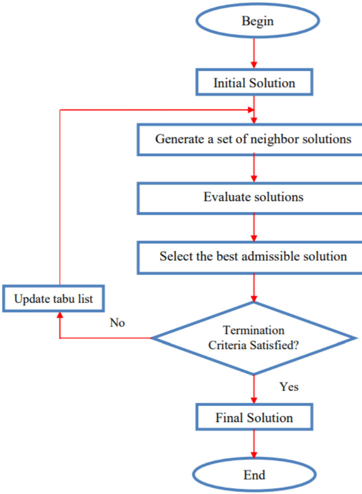

Figure 2: Tabu Search (Ali, 2016)

Genetic Algorithm

Yarkoni's (2010) implementation of GA is a metaheuristic that mimics evolutionary processes to search for solutions. A randomly selected groups of items making up potential solutions (chromosomes) make up the initial population, which is then evaluated with the fit function

$$Cost = Ik + \Sigma_{i=1}^{s}(1 - R_i^2)$$

where I is a fixed item cost, k is the number of retained items, i indexes the scale, and R_i^2 is the variance explained by the retained items on scale i. Smaller costs results in higher fitness; the more fit solutions are retained for the next iteration, with some leeway for crossover (two solutions trading sets of items) and mutations (random items switched out for others).

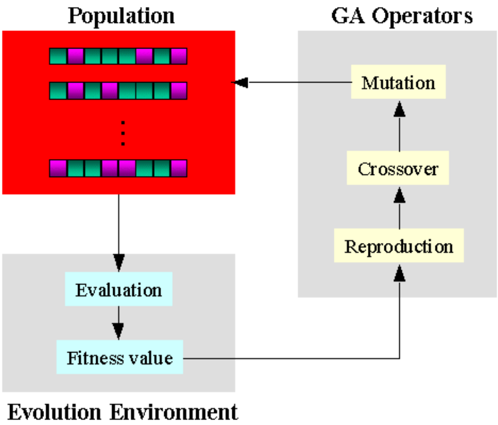

Figure 3: Genetic Algorithm (Liao & Sun, 2001)

Simulated Annealing

Simulated Annealing[3] (SA; Kirkpatrick, Gelatt, & Vecchi, 1983) is a global search metaheuristic that is a statistical analog to the metallurgic processes of annealing metals (Marcoulides & Drezner, 1999). It randomly searches the solution space and probabilistically accepts proposed solutions based on (1) the fit of the proposed solution ($model_2$) and (2) the temperature of the current iteration. The acceptance probability is

$$P(model_2 | fit_1, fit_2, currentTemp) = \begin{cases} exp\frac{-(fit_2 - fit_1)}{currentTemp}, & fit_1 > fit_2 \\ 1, & fit_1 \leq fit2 \end{cases}.$$

As the algorithm progresses, the temperature approaches zero, reducing the probability that worse-fitting solutions are selected.

[3] SA has not been used for psychometric models before in the literature.

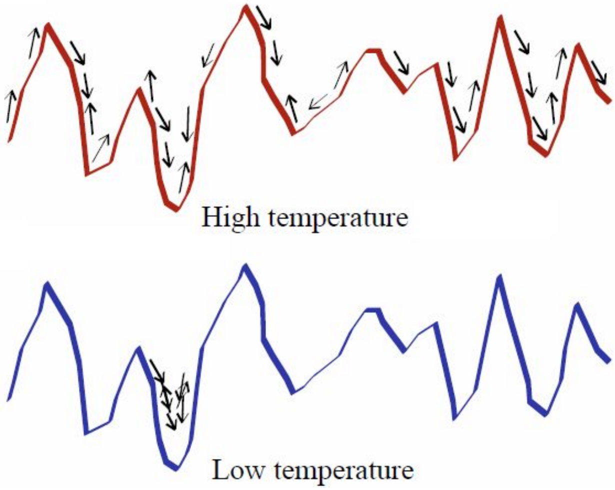

Figure 4: Simulated Annealing (Wang, 2013)

Research Questions

Research Questions

1. How do model misspecifications in the full form affect the fit of the short forms created by the algorithms?
2. Do the algorithms differ in their ability to exclude problematic items from the short forms?
3. Does the inclusion of a covariate (such as a predictive covariate or a convergent validity variable) affect the model fit of the short forms and the exclusion of problematic items?
4. How do the algorithms differ in terms of the time it takes for each to converge on a short form?

Method

Factors Manipulated

The following factors were manipulated for this study:

1. The dimensionality of the full form
   - One Factor
   - Three Factor
   - Bifactor with Three Specific Factors
2. Full-scale model misspecification
   - No misspecification
   - Minor misspecification (six items loading on a nuisance parameter with $\lambda = .3$)
   - Major misspecification (six items loading on a nuisance parameter with $\lambda = .6$)

3. Relationship to External Criterion Variable
   - No relationship
   - Moderate relationship ($\gamma = .6$)

This leads to a total of $3 * 3 * 2 = 18$ conditions in the study.

One Factor Model

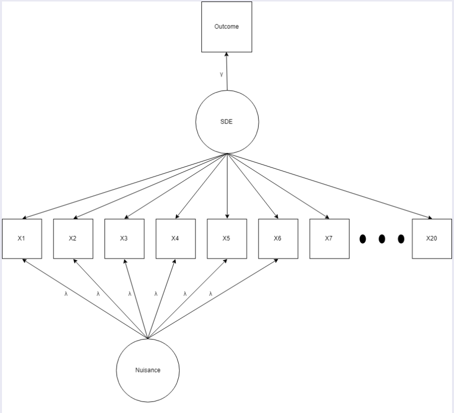

Figure 5: 20-item Self-Deceptive Enhancement Scale (Leite & Beretvas, 2005)

Three Factor Model

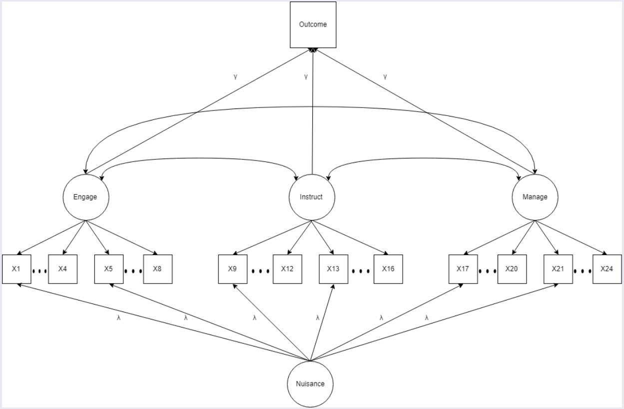

Figure 6: 24-item Teacher Efficacy Scale (Tschannen-Moran & Hoy, 2001)

Bifactor Model

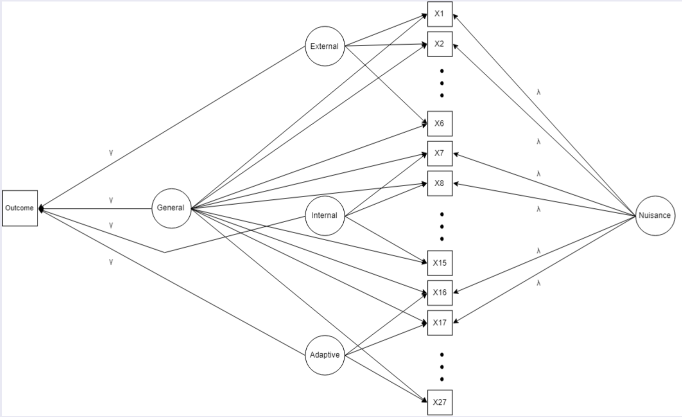

Figure 7: 27-item BASC-2 BESS Bifactor Model (Splett, Raborn, Lane, Binney, & Chafouleas, 2017)

Simulation

The entire analysis was conducted in R (R Core Team, 2018), with the data simulated with the `MASS` package (Venables & Ripley, 2002) using the covariance matrices for each condition.

The ACO, SA, and TS algorithm implementations used the `ShortForm` package (Raborn & Leite, 2018). The GA was adopted with minor modifications from the `GAabbreviate` package (Sahdra et al., 2016).

The sample size was fixed at $n = 500$, and a total of 100 iterations for each condition.

Analysis of Results

After data were simulated, the run time of each algorithm was recorded. Once the algorithms converged, the CFI, TLI, and RMSEA of the final model were saved and the composite reliability calculated using the following formula:

$$CR = \frac{\Sigma_{i=1}^{I} \lambda_i^2}{\Sigma_{i=1}^{I} \lambda_i^2 + \Sigma_{i=1}^{I} \theta_i}$$

where I is the total number of items in the short form, λ_i is the standardized factor loading of item i, and θ_i is the residual variance of item i. Finally, the proportion of iterations in which each algorithm included the problematic items was calculated.

Results

One Factor Model Fit: No External Variable

Error Condition	Method	Time to Complete (mins)	CFI	TLI	RMSEA	Composite Reliability
None	ACO	1.016	0.975	0.968	0.044	0.856
	SA	1.466	0.998	0.999	0.008	0.816
	Tabu	5.296	0.984	0.979	0.029	0.807
	GA	0.444	0.974	0.967	0.043	0.836
Minor	ACO	1.513	0.961	0.950	0.055	0.853
	SA	1.388	0.988	0.984	0.025	0.803
	Tabu	5.708	0.977	0.970	0.035	0.803
	GA	0.454	0.965	0.954	0.050	0.836
Major	ACO	1.248	0.902	0.874	0.095	0.853
	SA	1.154	0.983	0.978	0.030	0.789
	Tabu	4.433	0.941	0.924	0.062	0.806
	GA	0.448	0.846	0.801	0.113	0.834

One Factor Model Fit: External Variable

Error Condition	External Relationship	Method	Time to Complete (mins)	CFI	TLI	RMSEA	Composite Reliability
None							
	None	ACO	1.016	0.975	0.968	0.044	0.856
	None	SA	1.466	0.998	0.999	0.008	0.816
	None	TS	5.296	0.984	0.979	0.029	0.807
	None	GA	0.444	0.974	0.967	0.043	0.836
	Moderate	ACO	1.694	0.976	0.970	0.042	0.860
	Moderate	SA	1.552	0.991	0.990	0.020	0.812
	Moderate	TS	6.437	0.985	0.982	0.025	0.799
	Moderate	GA	0.448	0.977	0.971	0.041	0.858
Major							
	None	ACO	1.248	0.902	0.874	0.095	0.853
	None	SA	1.154	0.983	0.978	0.030	0.789
	None	TS	4.433	0.941	0.924	0.062	0.806
	None	GA	0.448	0.846	0.801	0.113	0.834
	Moderate	ACO	1.785	0.940	0.925	0.060	0.831
	Moderate	SA	1.574	0.982	0.977	0.028	0.782
	Moderate	TS	6.432	0.934	0.917	0.060	0.798
	Moderate	GA	0.449	0.856	0.819	0.109	0.852

One Factor Item Selection

Error Condition	Item	Factor Loading	ACO	SA	TS	GA
Minor	y3	0.58	81	42	32	56
	y5	0.534	87	41	47	43
	y4	0.448	70	57	60	58
	y2	0.408	76	41	45	44
	y6	0.393	32	25	42	90
	y1	0.382	43	42	49	50
Major	y3	0.58	76	20	52	59
	y5	0.534	87	25	46	41
	y4	0.448	72	18	60	78
	y2	0.408	66	27	51	43
	y6	0.393	65	17	41	93
	y1	0.382	70	19	44	53
Minor Error Proportion:			0.648	0.413	0.458	0.568
Major Error Proportion:			0.727	0.21	0.49	0.612

Three Factor Model Fit: No External Variable

Anthony Raborn[1], Walter Leite, Katerina Marcoulides

Error Condition	Method	Time to Complete (mins)	CFI	TLI	RMSEA	Composite Reliability
None	ACO	1.422	0.980	0.974	0.043	0.934
	SA	3.053	0.999	0.999	0.007	0.922
	Tabu	2.691	0.989	0.986	0.022	0.923
	GA	1.596	0.980	0.975	0.040	0.928
Minor	ACO	1.541	0.973	0.965	0.049	0.932
	SA	2.929	0.990	0.987	0.026	0.922
	Tabu	2.900	0.983	0.978	0.035	0.921
	GA	1.619	0.971	0.962	0.050	0.927
Major	ACO	1.307	0.949	0.935	0.064	0.929
	SA	2.413	0.990	0.987	0.027	0.921
	Tabu	2.310	0.913	0.888	0.079	0.919
	GA	1.596	0.907	0.880	0.089	0.922

Three Factor Model Fit: External Variable

Error Condition	External Relationship	Method	Time to Complete (mins)	CFI	TLI	RMSEA	Composite Reliability
None							
	None	ACO	1.422	0.980	0.974	0.043	0.934
	None	SA	3.053	0.999	0.999	0.007	0.922
	None	Tabu	2.691	0.989	0.986	0.022	0.923
	None	GA	1.596	0.980	0.975	0.040	0.928
	Moderate	ACO	1.770	0.978	0.971	0.046	0.936
	Moderate	SA	3.352	0.988	0.985	0.031	0.926
	Moderate	Tabu	3.397	0.983	0.978	0.038	0.924
	Moderate	GA	1.619	0.977	0.970	0.046	0.929
Major							
	None	ACO	1.307	0.949	0.935	0.064	0.929
	None	SA	2.413	0.990	0.987	0.027	0.921
	None	Tabu	2.310	0.913	0.888	0.079	0.919
	None	GA	1.596	0.907	0.880	0.089	0.922
	Moderate	ACO	1.894	0.960	0.948	0.059	0.930
	Moderate	SA	3.396	0.985	0.980	0.035	0.922
	Moderate	Tabu	3.412	0.952	0.938	0.062	0.921
	Moderate	GA	1.618	0.919	0.895	0.085	0.922

Three Factor Item Selection

Anthony Raborn[1], Walter Leite, Katerina Marcoulides

Error Condition	Item	Factor Loading	ACO	SA	TS	GA
Minor	y1	0.9	75	46	58	92
	y5	0.7	17	50	47	20
	y9	0.9	52	32	37	93
	y13	0.7	25	35	32	5
	y17	0.9	71	26	36	75
	y21	0.7	39	32	52	40
Major	y1	0.9	62	24	48	87
	y5	0.7	25	24	45	14
	y9	0.9	23	13	22	97
	y13	0.7	19	6	29	3
	y17	0.9	49	5	30	59
	y21	0.7	31	16	37	42
Minor Error Proportion:			0.465	0.368	0.437	0.542
Major Error Proportion:			0.348	0.147	0.352	0.503

Corresponding Author

anthony.w.raborn@gmail.com

References

References I

Ali, Z. A. (2016). *Concentric tabu search algorithm for solving traveling salesman problem* (Master's thesis). Eastern Mediterranean University (EMU)-Doğu Akdeniz Üniversitesi (DAÜ).

Byrne, G. J., & Pachana, N. A. (2011). Development and validation of a short form of the geriatric anxiety inventory–the gai-sf. *International Psychogeriatrics*, *23*(1), 125–131.

Colorni, A., Dorigo, M., Maniezzo, V., & others. (1992). An investigation of some properties of an" ant algorithm". In *PPSN* (Vol. 92, pp. 509–520).

Dennis, C.-L. (2003). The breastfeeding self-efficacy scale: Psychometric assessment of the short form. *Journal of Obstetric, Gynecologic, & Neonatal Nursing*, *32*(6), 734–744.

Dréo, J., Pétrowski, A., Siarry, P., & Taillard, E. (2006). *Metaheuristics for hard optimization: Methods and case studies*. Springer Science & Business Media.

Glover, F. (1989). Tabu search—part i. *ORSA Journal on Computing*, *1*(3), 190–206. http://doi.org/10.1287/ijoc.1.3.190

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*(4598), 671–680. http://doi.org/10.1126/science.220.4598.671

References II

Leite, W. L., & Beretvas, S. N. (2005). Validation of scores on the marlowe-crowne social desirability scale and the balanced inventory of desirable responding. *Educational and Psychological Measurement*, *65*(1), 140–154.

Leite, W. L., Huang, I.-C., & Marcoulides, G. A. (2008). Item selection for the development of short forms of scales using an ant colony optimization algorithm. *Multivariate Behavioral Research*, *43*(3), 411–431.

Liao, Y.-H., & Sun, C.-T. (2001). An educational genetic algorithms learning tool. *IEEE Transactions on Education*, *44*(2), 20–pp.

Lim, S. Y., & Chapman, E. (2013). Development of a short form of the attitudes toward mathematics inventory. *Educational Studies in Mathematics*, *82*(1), 145–164.

Marcoulides, G. A., & Drezner, Z. (2004). Tabu search variable selection with resource constraints. *Communications in Statistics-Simulation and Computation*, *33*(2), 355–362.

Marcoulides, G., & Drezner, Z. (1999). Using simulated annealing for model selection in multiple regression analysis. *Multiple Linear Regression Viewpoints*, *25*, 1–4.

Noble, W., Jensen, N. S., Naylor, G., Bhullar, N., & Akeroyd, M. A. (2013). A short form of the speech, spatial and qualities of hearing scale suitable for clinical use: The ssq12. *International Journal of Audiology*, *52*(6), 409–412.

Olaru, G., Witthöft, M., & Wilhelm, O. (2015). Methods matter: Testing competing models for designing short-scale big-five assessments. *Journal of Research in Personality*, *59*, 56–68.

Petrillo, G., Capone, V., Caso, D., & Keyes, C. L. (2015). The mental health continuum–short form (mhc–sf) as a measure of well-being in the italian context. *Social Indicators Research*, *121*(1), 291–312.

References III

Raborn, A., & Leite, W. (2018). *ShortForm: Automatic short form creation*. Retrieved from https://github.com/AnthonyRaborn/ShortForm

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Sahdra, K., B., Ciarrochi, J., Parker, P., . . . L. (2016). Using genetic algorithms in a large nationally representative american sample to abbreviate the Multidimensional Experiential Avoidance Questionnaire. *Frontiers in Psychology*, *7*(189), 1–14. Retrieved from http://www.frontiersin.org/quantitative_psychology_and_measurement/10.3389/fpsyg.2016.00189/abstract

Splett, J. W., Raborn, A., Lane, K. L., Binney, A. J., & Chafouleas, S. M. (2017). Factor analytic replication and model comparison of the basc-2 behavioral and emotional screening system. *Psychological Assessment*, *29*(12), 1543.

Toksari, M. D. (2016). A hybrid algorithm of ant colony optimization (aco) and iterated local search (ils) for estimating electricity domestic consumption: Case of turkey. *International Journal of Electrical Power & Energy Systems*, *78*, 776–782.

Tschannen-Moran, M., & Hoy, A. W. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education*, *17*(7), 783–805.

Veale, J. F. (2014). Edinburgh handedness inventory–short form: A revised version based on confirmatory factor analysis. *Laterality: Asymmetries of Body, Brain and Cognition*, *19*(2), 164–177.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth). New York: Springer. Retrieved from http://www.stats.ox.ac.uk/pub/MASS4

References IV

Wang, E. (2013). *A d3 plug-in for automatic label placement using simulated annealing*. Berkeley, Tech. Rep., Published as coursenotes: CS294-10-fa13.

Wester, S. R., Vogel, D. L., O'neil, J. M., & Danforth, L. (2012). Development and evaluation of the gender role conflict scale short form (grcs-sf). *Psychology of Men & Masculinity*, *13*(2), 199.

Yarkoni, T. (2010). The abbreviation of personality, or how to measure 200 personality scales with 200 items. *Journal of Research in Personality*, *44*(2), 180–198.