Unique Variable Analysis: A Novel Approach for Detecting Redundant Variables in Multivariate Data

Alexander P. Christensen[1], Luis Eduardo Garrido[2], and & Hudson Golino[3]

[1] University of Pennsylvania

[2] Pontificia Universidad Católica Madre y Maestra

[3] University of Virginia

Author Note

Acknowledgements: The authors did not preregister the study. All data, code, and materials can be found on the Open Science Framework: https://osf.io/9w3jy/. Alexander P. Christensen https://orcid.org/0000-0002-9798-7037, Luis Eduardo Garrido https://orcid.org/0000-0001-8932-6063, and Hudson Golino https://orcid.org/0000-0002-1601-1447.

The authors made the following contributions. Alexander P. Christensen: Conceptualization, Data Curation, Formal Analysis, Methodology, Software, Visualization, Writing - Original Draft, Writing - Review & Editing; Luis Eduardo Garrido: Conceptualization, Formal Analysis, Supervision, Writing - Review & Editing; Hudson Golino: Conceptualization, Formal Analysis, Data Curation, Supervision, Writing - Review & Editing.

Correspondence concerning this article should be addressed to Alexander P. Christensen, Department of Neurology, University of Pennsylvania, Philadelphia, PA, 19104. E-mail: alexpaulchristensen@gmail.com

Abstract

One common approach for constructing tests that measure a single attribute is the semantic similarity approach where items vary slightly in their wording and content. Despite being an effective strategy for ensuring high internal consistency, the information in tests may become redundant or worse confound the interpretation of the test scores. With the advent of network models, where tests represent a complex system and components (usually items) represent causally autonomous features, redundant variables may have inadvertent effects on the interpretation of their metrics. These issues motivated the development of a novel approach called *Unique Variable Analysis* (UVA), which detects redundant variables in multivariate data. The goal of UVA is to statistically identify potential redundancies in multivariate data so that researchers can make decisions about how best to handle them. Using a Monte Carlo simulation approach, we generated multivariate data with redundancies that were based on examples of known real-world redundancies. We then demonstrate the effects that redundancy can have on the accurate estimation of dimensions. Next, we evaluated UVA's ability to detect redundant variables in the simulated data. Based on these results, we provide a tutorial for how to apply UVA to real-world data. Our example data demonstrate that redundant variables create inaccurate estimates of dimensional structure but after applying UVA, the expected structure can be recovered. In sum, our study suggests that redundancy can have substantial effects on validity if left unchecked and that redundancy assessment should be integrated into standard validation practices.

Keywords: redundancy, minor factors, validity

Word count: 10,074 (main text), 244 (abstract)

Unique Variable Analysis: A Novel Approach for Detecting Redundant Variables in Multivariate Data

One of the core tenets of psychometrics is that a test should measure a single attribute. A primary strategy for developing tests that measure a single attribute is to select variables that are highly related to one another. This is reflected in traditional scale development conventions where variables are selected based on high item–test correlations and their contribution to the test's internal consistency (or the extent to which variables are interrelated; DeVellis, 2017). More recently, the advent of short forms and brief measures has renewed the focus on internal consistency as crucial aspect of test construction with researchers trying to maintain adequate levels of test information with fewer indicators (Clark & Watson, 2019). Indeed, internal consistency (Cronbach's α; Cronbach, 1951) is the most widely applied measure for item selection, and is often used (incorrectly) to determine whether a test measures a single attribute (Flake, Pek, & Hehman, 2017).[1]

The emphasis on internal consistency has led many to recommend dropping items with low internal consistency as a common step in scale development (Boateng, Neilands, Frongillo, Melgar-Quiñonez, & Young, 2018; MacKenzie, Podsakoff, & Podsakoff, 2011). On the one hand, selecting variables with high internal consistency increases the likelihood that a single attribute is being measured (DeVellis, 2017; McDonald, 1999). On the other hand, it can lead to redundancy between some variables in the test (Boyle, 1991; Reise, Bonifay, & Haviland, 2018). Some researchers have suggested the development of four-items scales should be summed up to form a "homogeneous scale," implicitly suggesting the use of redundant items (Comrey, 1988). Others have been critical of this approach call these repetitive scales "bloated specifics" (Cattell, 1978) and "cheating by repeating" (Reise, Bonifay, & Haviland, 2018, p. 681).

[1]We use internal consistency synonymously with Cronbach's α but note that there are other internal consistency measures such as McDonald's ω (McDonald, 1999) and other variants (McNeish, 2018) that are often better choices than Cronbach's α (especially when there are minor factors). We refer mainly to Cronbach's α in our discussion because it is still the most widely used internal consistency measure.

What is Redundancy?

We define *redundancy* broadly as large correlated uniqueness between latent variable indicators. This definition is congruent with local dependence in the item response theory (IRT) framework where variables are said to be locally dependent when they are significantly related after conditioning on the latent variable(s) in the model (McDonald, 1999). Locally dependent items contain less information than the IRT model would predict, meaning they possess redundant information (W.-H. Chen & Thissen, 1997). Some of the most common causes for item redundancy are a shared semantic reference (e.g., similar item phrasing, similar item content; Rosenbusch, Wanders, & Pit, 2020), shared substantive cause (e.g., single attribute), and general response tendency (e.g., social desirability; see Leising et al., 2020 for other potential causes).

Consistent with IRT, recent psychometric perspectives view redundancy less favorably particularly in areas where attributes are broad such as personality (A. P. Christensen, Golino, & Silvia, 2020; Clark & Watson, 2019; Cooper, 2019; McCrae & Mõttus, 2019; Saucier & Iurino, 2020). Researchers have placed increasing emphasis on global and local levels of it in their latent variable models (i.e., factor and IRT; Edelen & Reeve, 2007; Edwards, Houts, & Cai, 2018; Saris, Satorra, & Veld, 2009). More recently, psychometric network models have become concerned with redundant items because they affect the interpretation of their network measures (Hallquist, Wright, & Molenaar, 2019).

Psychometric network models represent a test as an interconnected network (Cramer, 2012; Epskamp, Maris, Waldrop, & Borsboom, 2018). In these networks, nodes (circles) represent variables and edges (lines) represent conditionally dependent relations (partial correlations) between the nodes (Epskamp & Fried, 2018). From this representation, network proponents advocate that attributes are represented as a complex system where components of these systems are defined as *causally autonomous* or having distinct causal processes (Cramer et al., 2012). Such a perspective implies that *network components*—variables or sets of variables that share a unique common cause—should *not* be redundant (A. P. Christensen, Golino, & Silvia, 2020).

Present Research

Item redundancy is most problematic in exploratory research where the latent structure of a set of variables has yet to be established. In this context, one of the crucial decisions faced by researchers is the determination of the number of major factors (Garrido, Abad, & Ponsoda, 2013; Hayton, Allen, & Scarpello, 2004; Henson & Roberts, 2006). While several accurate techniques are available to aid in this process (Golino et al., 2020), their estimates are impacted by item redundancy, leading to an overestimation of the number of major dimensions and the potential selection of sub-optimal latent structures (A. P. Christensen, Gross, Golino, Silvia, & Kwapil, 2019; Flores-Kanter, Garrido, Moretti, & Medrano, 2021). In order to combat this issue, researchers may try to identify and address item redundancy prior to performing a dimensionality assessment of the data. Unfortunately, the measures to detect item redundancy in both factor analysis (i.e., residual correlations) and IRT (i.e., local dependence statistics; Yen, 1984) require the specification of the latent structure of the data, something that is not available in the exploratory stages of scale validation.

To address this problem we developed a new procedure called *Unique Variable Analysis* (UVA) that is based on psychometric network analysis and that does not require *a priori* knowledge of the latent structure of the data. first conducted a simulation study to assess the impact of item redundancy on dimensionality assessment, and then evaluated UVA's ability to detect item redundancy across a wide range of multidimensional structures. As a comparison, we also examined a widely used and recommended IRT local dependence measure: Yen's Q3 statistic (Yen, 1984). We organize the rest of the paper by giving a brief overview on the potential problems of item redundancy, describing the proposed UVA network approach to detect item redundancy, presenting the objectives of the current study, reporting the results of our simulation study, and demonstrating how UVA can be applied to real-world data.

Potential Problems of Item Redundancy

A recent simulation study illuminated one psychometric consequence that redundancy can have on the interpretation of network measures. Hallquist, Wright, and Molenaar (2019) performed a set of three simulation studies to evaluate the effects of multiple latent causes on commonly used network measures called *centrality*. Centrality measures quantify the relative position of a node in a network. The three most frequently used measures have been *betweenness* (nodes used most often on the shortest path between other nodes), *closeness* (nodes with the shortest path on average to other nodes), and *strength* (absolute sum of a node's connections). In their simulations, Hallquist and colleagues (2019) generated data from common factors models, which allowed them to simulate multiple latent causes on specific variables. This approach enabled them to evaluate the effect that multiple latent causes may have on the psychometric interpretation of centrality measures.

Their most relevant simulation generated data with two factors and ten variables per factor. One variable on each factor, which we refer to as *target variables*, also loaded onto a third factor. In practical terms, such a factor might represent shared semantic similarities between two variables (e.g., "I like to go to parties" and "I like to go out with people"; Leising et al., 2020; Rosenbusch, Wanders, & Pit, 2020). They incrementally manipulated the loadings of the target variables between the shared and non-shared factors, forming a gradient of shared to non-shared factor effects on the centrality measures. They found that all measures were nonlinearly affected by the presence of multiple causes on a variable, which made centrality difficult to interpret. Because of this finding, Hallquist et al. (2019) recommended that "new measures be developed using methods that capture [causally distinct] components while discouraging the presence of latent confounding" (p. 19). In other words, redundancy broadly should be reduced in psychometric network models to more accurately interpret centrality measures.

Network measures may not be all that is affected. Dimensionality, for instance, is likely affected if there are strong enough relations between subsets of variables to form *minor factors* or factors of correlated residual variance (or unique variance; Garrido et al., 2020). In these cases, dimensionality assessment methods may suggest more dimensions than the number of major factors present in the population. This overidentification is concerning because the number of factors to extract is a critical decision when conducting exploratory factor analysis (EFA) or exploratory structural equation modeling (ESEM; Garrido, Abad, & Ponsoda, 2013; Hayton, Allen, & Scarpello, 2004; Henson & Roberts, 2006). When more dimensions are estimated in these exploratory factor methods than the number of major factors in the population, the recovery of the major factors deteriorates (Fava & Velicer, 1992). One potential issue is known as "factor splitting" or where groups of variables from a major factor separate and load on different (new) factors (Wood, Tataryn, & Gorsuch, 1996). This issue is especially problematic because the split factors could be (and often are by applied researchers) interpreted as distinct entities, and theoretically developments could be undertaken to clarify their supposed differences.

One example of factor splitting comes from work which examined the Positive and Negative Affective Schedule (PANAS; Watson, Clark, & Tellegen, 1988). The PANAS is comprised of two theoretical dimensions: Positive Affect and Negative Affect. However, it was recently shown that item redundancies led to three dimensions estimated by exploratory graph analysis [EGA; Golino and Epskamp (2017); Golino et al. (2020)], a network psychometrics approach for factor estimation (Flores-Kanter, Garrido, Moretti, & Medrano, 2021). The Negative Affect factor was splitting into two factors; additionally, the dimensionality estimates were very unstable, fluctuating between two and three factors across bootstrap samples. In contrast, when the item redundancies were addressed, EGA suggested the theoretical two-factor structure and the estimates were completely stable.

Psychometric network models are not the only models affected by redundancy though. Redundancy can have similar consequences for latent variable models as well (e.g., factor analysis, IRT; Embretson & Reise, 2000). Redundancy, regardless of how or why, is a violation of the local independence assumption that often leads to significant correlated residuals, often resulting in poor fitting models (e.g., Big Five personality traits; Hopwood & Donnellan, 2010; Montoya & Edwards, 2020) and confounding the interpretation of test scores (Gerbing & Anderson, 1984; Goldstein, 1980). Although strict adherence to local independence is likely unavoidable in psychology, its violations go largely unchecked in most validation research (Flake, Pek, & Hehman, 2017).

A Novel Approach to Detect Redundancy

Factor analytic approaches for determining whether variables are redundant require estimating a latent variable model first and then check the correlations between the residuals (Saris, Satorra, & Veld, 2009). Limitations of estimating a latent variable model first is that the correlated errors will change depending on the model (e.g., number of factors specified), conflating evidence of redundancy. In the case of IRT, numerous measures have been developed to detect local dependence, with the two most well-known indices being Yen's Q3 (Yen, 1984) and the likelihood ratio statistic G^2 (W.-H. Chen & Thissen, 1997; see Edwards, Houts, & Cai, 2018 for more details). These measures, however, require knowledge of the number and composition of the major factors.

Knowledge of the number and composition of major factors is often not known in exploratory research, leaving researchers to either assume a structure or apply these measures to their entire dataset. Unfortunately, such an approach would likely lead to poor results because the local dependence measures would conflate item redundancy with substantive multidimensionality (W.-H. Chen & Thissen, 1997). To overcome these limitations, we developed a novel approach that does not require the estimation of latent factors or prior knowledge of the latent structure.

Our approach uses psychometric network analysis to first estimate a network. To estimate a network, we used the standard in the psychometric network literature: the graphical least absolute shrinkage and selection operator [GLASSO; Epskamp and Fried (2018); Friedman, Hastie, and Tibshirani (2008); Friedman, Hastie, and Tibshirani (2014)], following the approach used in EGA (see SI4; Golino et al., 2020). After estimating a network, the network measure called weighted topological overlap (wTO; Zhang & Horvath, 2005) is applied.

wTO measures the extent to which items in a network "overlap" by quantifying the similarity between a pair of variables' shared connections (e.g., weights, signs, quantity). These measures became popular in the biological sciences because of their ability to identify hierarchical organizations in metabolic and genetic networks (Nowick, Gernat, Almaas, & Stubbs, 2009; Ravasz, Somera, Mongru, Oltvai, & Barabási, 2002). The original topological overlap was derived for unweighted (binary) networks (equation from methods supplement of Ravasz et al., 2002 and amended by Zhang & Horvath, 2005):

$$\omega_{ij} = \frac{l_{ij} \ + \ a_{ij}}{\min\{k_i, k_j\} \ + \ 1 \ - \ a_{ij}},$$

where $l_{ij} = \sum_u a_{iu} a_{uj}$, $k_i = \sum_u a_{iu}$, and a_{ij} is the weighted between nodes i and j. In the psychometric network literature, k_i is often referred to as *degree* or the number of connections for node i. l_{ij} corresponds to the number of nodes that both i and j are connected to. This equation holds for weighted networks where k_i instead refers to node strength (or the sum of a node's connections) for node i. Our implementation of weighted topological overlap uses the *wTO* package (Gysi, Voigt, de Miranda Fragoso, Almaas, & Nowick, 2018) in R (R Core Team, 2020), which uses Nowick and colleagues' (2009) adjustment for signed values. For the above equation, this means changing $mink_i, k_j$ and a_{ij} in the denominator to be absolute values.

Our general approach works as follows: (1) estimate a network (see SI4), (2) compute weighted topological overlap between each variable pair, creating an overlap matrix, (2) obtain the lower triangle of the overlap matrix (to count weights once), (3) remove values that equal zero so that only non-zero values remain, and (4) based on the remaining absolute values apply a threshold. The thresholds applied are either a cut-off (e.g., .20) or percentile. If using a percentile approach, then an empirical distribution (e.g., gamma) is fit and p-values are obtained for each weight. Each weight in the overlap matrix corresponds to a pair of variables and the p-values correspond to percentiles (Perezgonzalez, 2015). The weights that survive an α threshold (e.g., $\leq .05$) correspond to two variables that are redundant.

Aims

We had three aims for the present research. First, we wanted to assess the impact of item redundancy on dimensionality estimation using the state-of-the-art network psychometrics approach of EGA. Previous research has examined and demonstrated that even a few correlated residuals can reduce model fit and lead to overfactoring in structural equation models (Montoya & Edwards, 2020). To our knowledge, effects of correlated residuals has not been explored using exploratory techniques. EGA tends to be at least as accurate and less biased as some of the best exploratory factor analytic methods (e.g., parallel analysis; A. P. Christensen, Garrido, & Golino, 2021; Golino et al., 2020); therefore, any effects that redundancy has on EGA's accuracy and bias would also be expected for any other factor analytic method. We performed a Monte Carlo simulation where we manipulated sample size, number of factors, number of variables per factor, correlations between factors, and number of redundant variables within each factor. To generate data with realistic redundancies, we identified known redundancies in real-world data and mimicked their characteristics in our data generation method. We then evaluated the performance of EGA by computing measures of accuracy and bias. Our hypothesis was that for structures with redundant items, EGA would suggest more dimensions than the number of major factors in the population.

Our second and primary aim was to evaluate the accuracy of the proposed UVA approach in the detection of redundant variables. We evaluated our approach by comparing it to a contemporary IRT standard: Yen's Q3 local dependence statistic (Yen, 1984). For both approaches, we used thresholds suggested in the literature and thresholds determined to have optimal accuracy based on a grid search. We also applied percentiles based on standard and adaptive α values (Pérez & Pericchi, 2014). Sensitivity and specificity measures were used to quantify the extent to which these method and threshold combinations were most accurate. Our hypothesis in this case was that UVA would be more accurate in the detection of item redundancy than Q3, as the latter measure was expected to confound redundancy in multidimensional structures (Boyle, 1991).

Based on our simulation results, our third aim was to demonstrate how UVA can be applied to real-world data. Using data with a theoretically grounded structure from the *Broad Autism Phenotype Questionnaire* (BAPQ; Hurley, Losh, Parlier, Reznick, & Piven, 2007), we show that redundancy can produce split factors that lead to an inaccurate estimation of the expected structure. We then provide a tutorial on how to apply an automated version of UVA to this data. Afterwards, we re-estimate the dimensionality of the data and show that the expected dimensional structure can be recovered.

Methods

Data Generation

Our data generation procedure followed the same approach as Golino et al. (2020), generating a r-dimensional item factor analysis model. First, the reproduced population correlation matrix was computed:

$$\mathbf{R_R} = \boldsymbol{\Lambda\Phi\Lambda}',$$

where $\mathbf{R_R}$ is the reproduced population correlation matrix, lambda ($\boldsymbol{\Lambda}$) is the k (variables) $\times$ r (factors) factor loading matrix, and $\boldsymbol{\Phi}$ is the $r \times r$ correlation matrix. The population correlation matrix, $\mathbf{R_P}$, was then obtained by putting the unities on the diagonal of $\mathbf{R_R}$. Next, Cholesky decomposition was performed on the correlation matrix such that:

$$\mathbf{R_P} = \mathbf{U}'\mathbf{U}.$$

If the population correlation matrix was not positive definite (i.e., at least one eigenvalue ≤ 0) or any single item's communality was greater than 0.90, then $\mathbf{\Lambda}$ was re-generated and the same procedure was followed until these criteria were met. Finally, the sample data matrix of continuous variables was computed:

$$\mathbf{X} = \mathbf{ZU},$$

where $\mathbf{Z}$ is a matrix of random multivariate normal data with rows equal to the sample size and columns equal to the number of variables. To generate polytomous data, each continuous variable was categorized with a skew of zero onto a 5-point Likert scale, mimicking many commonly used assessment instruments (Garrido, Abad, & Ponsoda, 2013). All data generated in this study were polytomous.

Simulation Design

Across population models, factor loadings for each variable were randomly drawn from a uniform distribution with values between .40 and .70 to mimic more realistic data conditions. Similarly, cross-loadings were generated following a random normal distribution with a mean of zero and a standard deviation of .10. This procedure follows previous simulation work described in Garcia-Garzón, Abad, and Garrido (2019). Including cross-loadings provides data that are more likely to be found in real-word data (Bollmann, Heene, Küchenhoff, & Bühner, 2015).

Two, three, and four factors were simulated to provide multidimensional structures that are commonly found in the psychological literature (Henson & Roberts, 2006). There were six and twelve variables per factor, which were chosen to evenly produce proportions of the redundant variables per factor: 0.000, 0.167, 0.333, and 0.500. The condition of zero variables was particularly important for estimating the consistency for which methods had false positives. Correlations between factors were manipulated to be orthogonal (0.00), small (0.30), moderate (0.50), and large (0.70). Finally, very small (250), small (500), and medium (1000) sample sizes were generated to mimic typical sample sizes in psychological psychometric work.

The simulation design allowed for a mixed factorial design: $3 \times 2 \times 4 \times 4 \times 3$ (number of factors $\times$ variables per factor $\times$ proportion of redundant items $\times$ correlations between factors $\times$ sample size) for a total of 288 simulated condition combinations. For condition, 500 samples were generated, totaling 144,000 samples across conditions.

Simulating Redundancy

Our approach for generating redundancies in the data closely followed the so-called "threshold-shift" formulation used in IRT (Glas & Falcón, 2003). In the threshold-shift formulation, one item is made to be dependent on one or more other items (see Eq. 13 in Glas & Falcón, 2003). We mimicked this procedure by replacing a subset of values in one variable with an equal subset from another variable. To avoid replacing exact values, random noise was added. The random noise, on average, added or subtracted one standard deviation from the copied value. This procedure was followed for each variable that was made "redundant."

To ensure that redundant variables were consistent with real-world redundancies in empirical data, we quantified real-world redundancies and calibrated our data generating parameters. We first obtained descriptive pairwise response pattern metrics of Synthetic Aperture Personality Assessment (SAPA; Condon, 2018; see also Revelle, Dworak, & Condon, 2020) items that demonstrated obvious and substantial overlap in either content or concepts of items (see SI1 for details about the SAPA inventory). We identified thirty-six pairs of items (see SI2 for all item labels and their descriptions, and SI3 for the redundant pairings). For these items, we computed pairwise response pattern metrics that quantified the proportion of exact matches between responses (PEM; a participant responded with "3" for both items) and the mean absolute error (MAE) or difference between responses on two items. If necessary, we reverse scored items so their item valences were in the same direction. The average PEM was 0.41 (*median* = 0.40, *SD* = 0.08, *range* = 0.22–0.60) and the average MAE was 0.93 (*median* = 0.91, *SD* = 0.24, *range* = 0.54–1.62).

For the proportion of values that were copied to replace values in the replace set, we calibrated our data generating procedure to attempt to match realistic conditions—that is, matching the observed PEM and MAE metrics mentioned above. To do so, we generated continuous data with 50% of values copied for each copy to replace variable, categorized them, and checked our simulated data's PEM and MAE against the observed data's PEM and MAE. This process continued—generating data and checking the PEM and MAE values—until we were satisfied with the result. Our simulated data achieved an average PEM of 0.38 (*median* = 0.37, *SD* = 0.08, *range* = 0.16–0.69) and an average MAE of 0.71 (*median* = 0.72, *SD* = 0.13, *range* = 0.32–1.24). We note that our MAE was smaller than the observed data, which suggests that the dispersion of response values for the simulated data were not as wide as the observed data.

Dimensionality Estimation

In traditional psychometric research, some redundancy has allowed scales to be developed that are aimed at targeting specific psychological attributes (DeVellis, 2017). For psychometric network models, however, this redundancy has some potential consequences such as marring the interpretability of network measures (Hallquist, Wright, & Molenaar, 2019) and obscuring dimensionality estimates due to subsets of variables and forming minor factors (e.g., A. P. Christensen, Gross, Golino, Silvia, & Kwapil, 2019).

To provide a demonstration of how redundancy affects dimensionality, we applied EGA (Golino & Epskamp, 2017; Golino et al., 2020) to our simulated data (see SI4 for technical details). Simulation studies have consistently demonstrated that EGA is as accurate as state-of-the-art factor analytic approaches, such parallel analysis, but also tends to be less biased (i.e., underfactoring or overfactoring; A. P. Christensen, Garrido, & Golino, 2021; Golino et al., 2020). Importantly, redundancy was not simulated in these studies and therefore remains an important condition in determining how dimensionality estimates are affected. Regardless of the dimensionality reduction method applied, redundancy is likely to affect estimates in the same way: overfactoring due to subsets of redundant variables forming minor factors (Montoya & Edwards, 2020).

Item Redundancy Detection

To evaluate whether variables were redundant, we evaluated the performance of our proposed UVA approach and Yen's Q3 IRT local dependence statistic. A unidimensional IRT model for was employed for Q3 to be consistent with the notion of not implying a specific factor structure. For threshold types, we used the standard alpha (α = .05; hereafter referred to as alpha), adaptive alpha (Pérez & Pericchi, 2014), and threshold. We note that alpha, conceptualized as a percentile (i.e., 95^{th}), has been used in combination with Q3 in the IRT literature (K. B. Christensen, Makransky, & Horton, 2017). Adaptive alpha was used as an alpha adjustment that was flexible based on the number of variable pairs (details of adaptive alpha can be found in the Supplementary Information; SI5).

For the thresholds, we used two approaches: standard and optimal. The standard threshold was based on previous recommendations from the literature. For Q3, we applied a threshold of .20, which is commonly used in the literature (Marais, 2012). For weighted topological overlap, we used .30 based on previous evidence reported by Nowick, Gernat, Almaas, and Stubbs (2009), which found values greater than .30 were not obtained when genes associated with the brain were permutated fifty times for each person in their sample. The optimal thresholds were based on a grid search applied to the simulated data. For both Q3 and weighted topological overlap, we computed accuracy metrics for each redundancy proportion across thresholds of .15 to .40 on .05 increments. We found that Q3 had the highest accuracy with a threshold of .35 and weighted topological overlap had the highest accuracy with a threshold of .25 (discussed further in our Results section). These thresholds were used in comparison to our other threshold types. IRT and Q3 were estimated and computed using the *mirt* package (Chalmers, 2012) in R.

Statistical Analyses

Effects of Redundancy on Dimensionality

We computed one measure of accuracy and one measure of bias for our dimensionality assessment. For accuracy, we computed percent correct (PC) or the average of the number of times EGA correctly estimated the population number of factors across conditions. For bias, we computed mean bias error (MBE) or the average difference between the population number of factors across conditions. The former directly captures the general accuracy of the estimation, while the latter captures whether EGA tends to underfactor or overfactor. Based on our simulation design, we expected that EGA would consistently overfactor due to the redundancies in the data. We computed analysis of variances (ANOVAs) with interactions between all conditions. Partial eta-squared (η^2) effect sizes were used to determine whether there was a main or interaction effect. We followed Cohen's (1992) effect size guidelines: small = .01, moderate = .06, and large = .14.

Evaluation of Redundancy Approach

To evaluate the performance of the redundancy approaches, we used sensitivity and specificity measures. These measures are based on true positives (TP; detecting redundancy when there was redundancy between pairs of variables), true negatives (TN; detecting no redundancy when there was no redundancy between pairs of variables), false positives (FP; detecting redundancy when there was no redundancy between pairs of variables), and false negatives (FN; detecting no redundancy when there was redundancy between pairs of variables). For the condition with no redundancies, we computed specificity ($\frac{TN}{TN+FP}$); for conditions with redundancies, we computed the critical success index (CSI; $\frac{TP}{TP+FP+FN}$). To simplify presentation, we used these measures as "accuracy" for their respective conditions.

Data Analysis

We used R (version 4.0.5; R Core Team, 2020) for all of our analyses and the *papaja* package (version 0.1.0.9997; Aust & Barth, 2020) for our manuscript preparation. The package *EGAnet* (version 0.9.9; Golino & Christensen, 2020) was used to apply EGA and UVA, and figures were created using *ggplot2* (version 3.3.3; Wickham, 2016) and *GGally* (version 2.1.0; Schloerke et al., 2020). The session information for all necessary R packages can be found in our Supplementary Information (SI6). All data and R scripts can be found on the Open Science Framework: https://osf.io/9w3jy/.

Results

Redundancy Effects on Expected Dimensionality

Our first goal was to examine the effects that redundancy can have on the expected dimensional structure of the data (i.e., the ground truth number of dimensions in the data before manipulating variables to be redundant). In applied terms, the expected structure corresponds to a theoretical structure (e.g., Big Five personality traits) that corresponds the structure of the major dimensions. Importantly, by adding redundancy to the data, we are manipulating the ground truth structure away from the expected structure—that is, we are creating the potential for additional (minor) factors in the data. In this way, the structure estimated by EGA could be correctly estimating the number of total dimensions in the data; however, the number of major dimensions is incorrect in terms of the expected structure because of the redundancy. Therefore, our intention was to determine the extent to which redundancy affects the estimation of the expected (major) dimensions.

We first examined the effect sizes of the conditions and their interactions using ANOVAs. For PC, this revealed that there was an overwhelmingly large effect size for the proportion of redundant variables ($\eta_p^2 = 0.55$). As is clear in Figure 1, redundancy had a substantial effect on EGA's accuracy to the extent that performance dropped from 97.6% with no redundancy (0.000) to 33.2% with minimal redundancy (0.167). When there was moderate (0.333) to large (0.500) redundancy, EGA was almost never estimating the proper number of factors (11.9% and 2.0%, respectively). Such negative effects on accuracy are beyond any condition examined so far in simulations using EGA (A. P. Christensen, Garrido, & Golino, 2021; Golino & Epskamp, 2017; Golino et al., 2020).

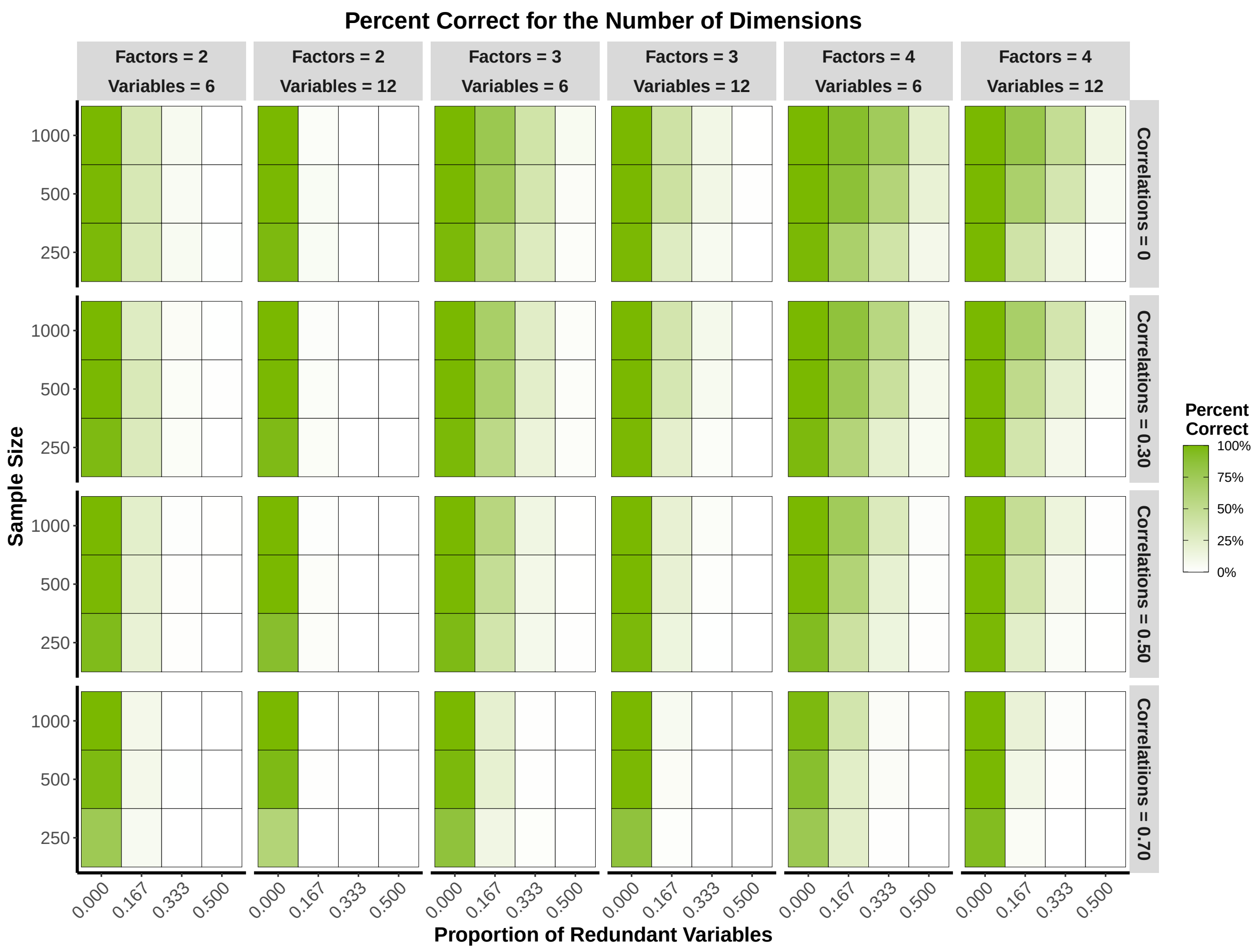

Figure 1

Redundancies Effect on the Accruacy of EGA

For MBE, there was a large main effect of proportion of redundant variables ($\eta_p^2 = 0.65$), number of variables per factor ($\eta_p^2 = 0.23$), and correlations between factors ($\eta_p^2 = 0.14$). There was a moderate-to-large effect size for the interaction between proportion of redundant variables and number of variables per factor ($\eta_p^2 = 0.13$). The interaction, however, occurred below the possible value of proportion of redundant variables (0.000; see SI7). We interpreted this interaction to suggest that as the proportion of variables increased the difference between MBE for the number of variables per factor increased with more variables having a sharper positive slope (i.e., more redundancy, the greater the increase of MBE for more variables per factors). For most conditions, if there was bias, then it was in the direction of overfactoring (MBE > 0) rather than underfactoring (MBE < 0; Figure 2).

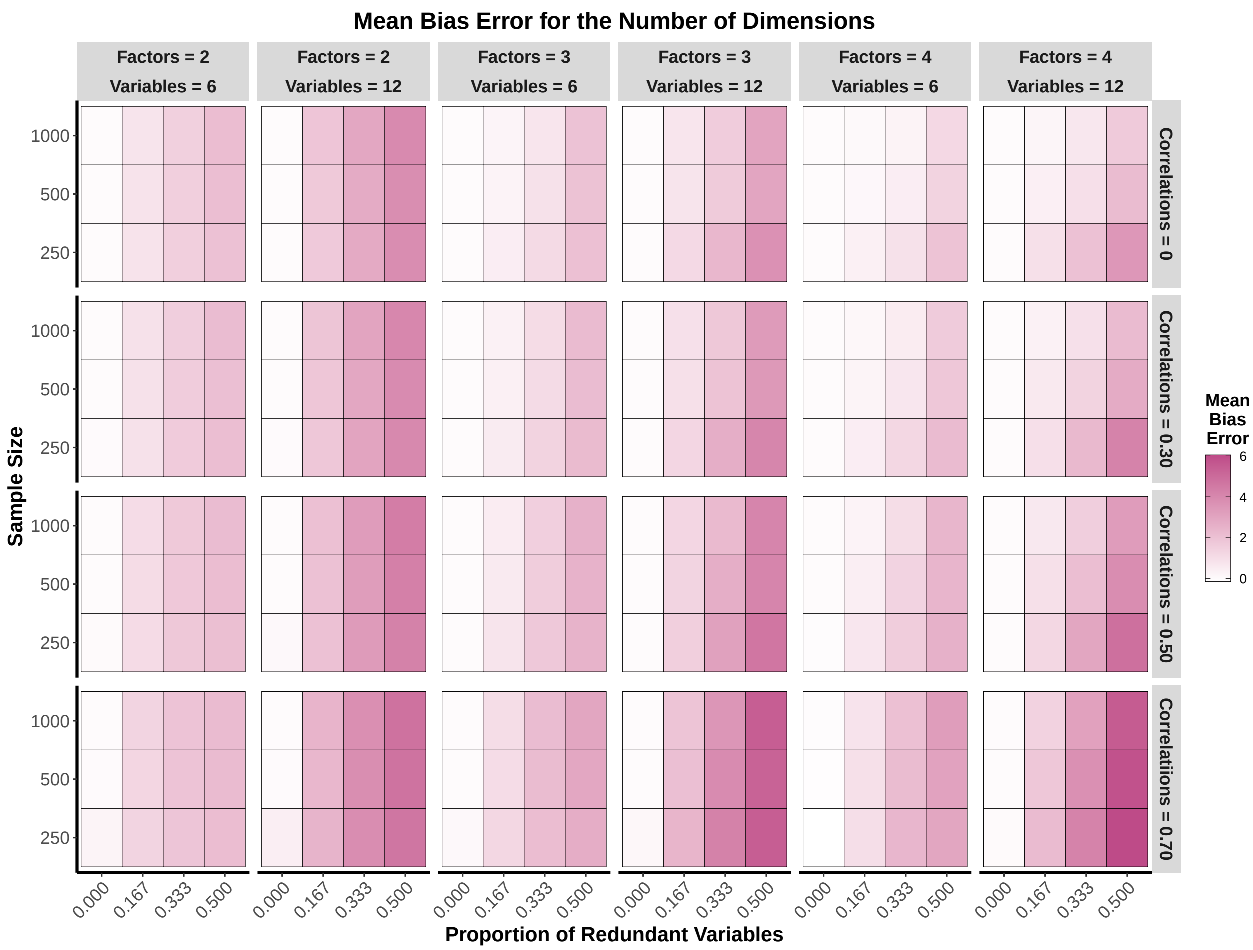

Figure 2

Redundancies Effect on the Mean Bias Error of EGA

In short, our results demonstrate that redundant variables have a strong effect on the accuracy and bias of dimensionality estimation of the expected factor structure with EGA. Redundancy interacted with number of variables per factor, suggesting that more variables and more redundancy will produce worse overfactoring bias.

Grid Search for Optimal Redundancy Thresholds

We performed a grid search to determine the optimal accuracy for the thresholds for Yen's Q3 of IRT and weighted topological overlap. Each threshold was applied to our simulated data and broken down by each proportion of redundant variables condition (Figure 3).

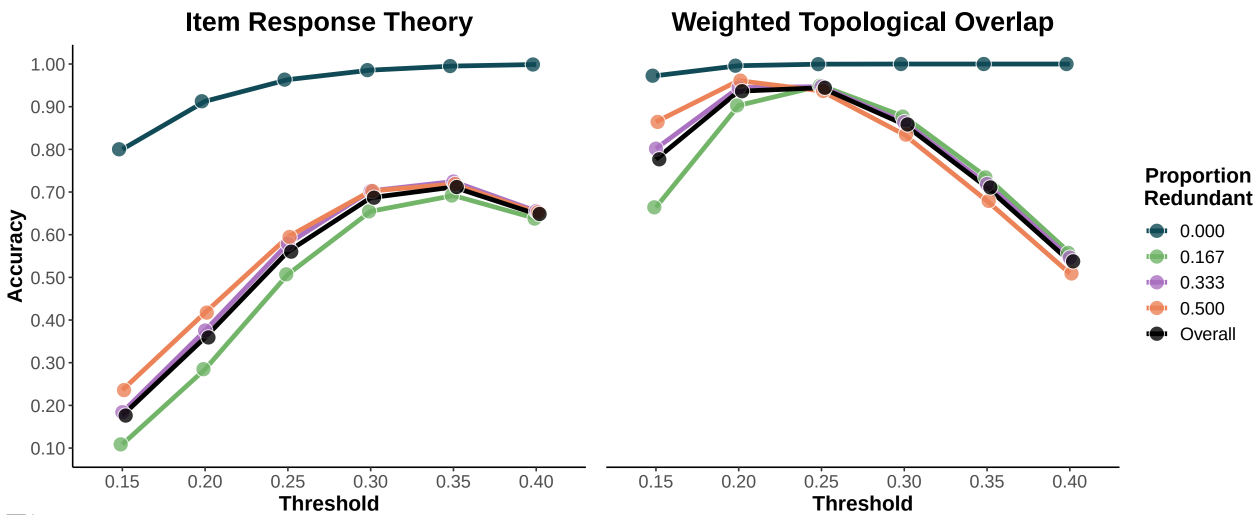

Figure 3

Grid search for IRT (left) and weighted topological overlap (right).

For IRT, accuracy increased as the threshold increased until reaching 0.35 where accuracy started to decrease. For no redundancies, the highest threshold in the search, 0.40, had the highest accuracy (0.999) followed by thresholds of 0.35 (0.995) and 0.30 (0.987). For redundancies, 0.35 had the highest accuracy (0.712) followed by 0.30 (0.689) and 0.40 (0.659; black line in Figure 3). For all non-zero redundancy conditions, the threshold of 0.35 had the highest accuracy. Despite the threshold of 0.20 being the common cut-off in the literature (W.-H. Chen & Thissen, 1997; Marais, 2012; Yen, 1984), it was among the least accurate threshold in our IRT grid search (0.359).

For weighted topological overlap, accuracy increased and plateaued around 0.20 and 0.25 where it decreased for every threshold thereafter. For no redundancies, the four highest thresholds had perfect accuracy (1.000). For redundancies, 0.25 had the highest accuracy (0.938) followed by 0.20 (0.929) and 0.30 (0.843; black line in Figure 3). Best performers varied based on the proportion of redundant variables. For the 0.167 proportion of redundant variables, 0.25 performed the best (0.944) followed by 0.20 (0.892) and 0.30 (0.863). For 0.333, 0.25 and 0.20 were comparable (0.941 and 0.939, respectively) followed by 0.30 (0.847). For 0.500, 0.20 performed the best (0.957) followed by 0.25 (0.930) and 0.15 (0.849). For all non-zero redundancy conditions, the threshold of 0.25 had the highest accuracy (0.938) followed by 0.20 (0.929). Despite the threshold of 0.30 being used in previous genetics research (Nowick, Gernat, Almaas, & Stubbs, 2009), it was the third best performing thresholds when there was redundancy in our psychological data (0.843).

Performance of the Redundancy Approach

IRT

We start first by breaking down the results by each method (i.e., IRT and weighted topological overlap) and then compare the overall results together. For IRT, there was a moderate interaction between the proportion of redundant variables and correlation between factors ($\eta_p^2 = 0.10$). Based on this interaction, we examined accuracy broken down by correlations between factors (Figure 4).

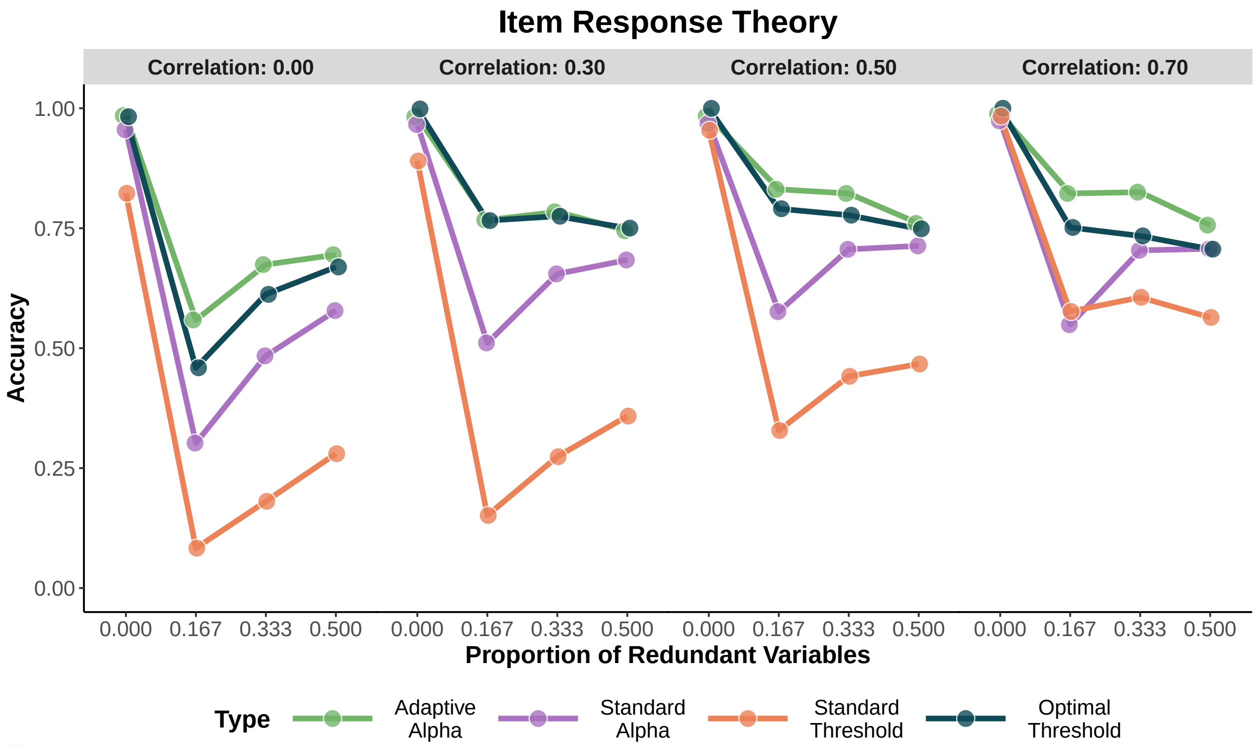

Figure 4

Accuracy of IRT and threshold types. Correlation = correlation between factors.

As the correlations between factors increased, accuracy for all IRT and threshold types also increased. Such a result is consistent with IRT's assumption of unidimensionality—correlations between factors in multidimensional structures are a proxy for unidimensionality violation (0.00 = stronger violation to 0.70 = weaker violation). Across the correlations between factors, the optimal threshold had the highest accuracy for the no redundancy conditions. For the 0.167 proportion of redundant variables, adaptive alpha had the highest accuracy for correlations between factors of 0.00 (0.559), 0.50 (0.831), and 0.70 (0.822). Adaptive alpha and optimal threshold had comparable accuracy for 0.30 (both = 0.767). For 0.333 and 0.500, a similar pattern held where adaptive alpha had the highest accuracy for correlations between factors of 0.00, 0.50, and 0.70. For correlations between factors of 0.30, adaptive alpha (0.784) was slightly higher than the optimal threshold (0.775) when the proportion of redundant variables was 0.333, while the optimal threshold (0.750) was slightly higher than adaptive alpha (0.745) when the proportion of redundant variables was 0.500. By far, the standard threshold had the worst accuracy across correlations between factors and proportion of redundant variables conditions. The standard threshold was only slightly better when than standard alpha when the correlations between factors were high (0.70) and the proportion of redundant variables were low (0.00 and 0.167).

Weighted Topological Overlap

For weighted topological overlap, there were two separate interactions for proportion of redundant variables: number of factors ($\eta_p^2 = 0.06$) and number of variables per factor ($\eta_p^2 = 0.07$). Based on these interaction, we examined accuracy broken down by number of factors and number of variables per factor (Figure 5).

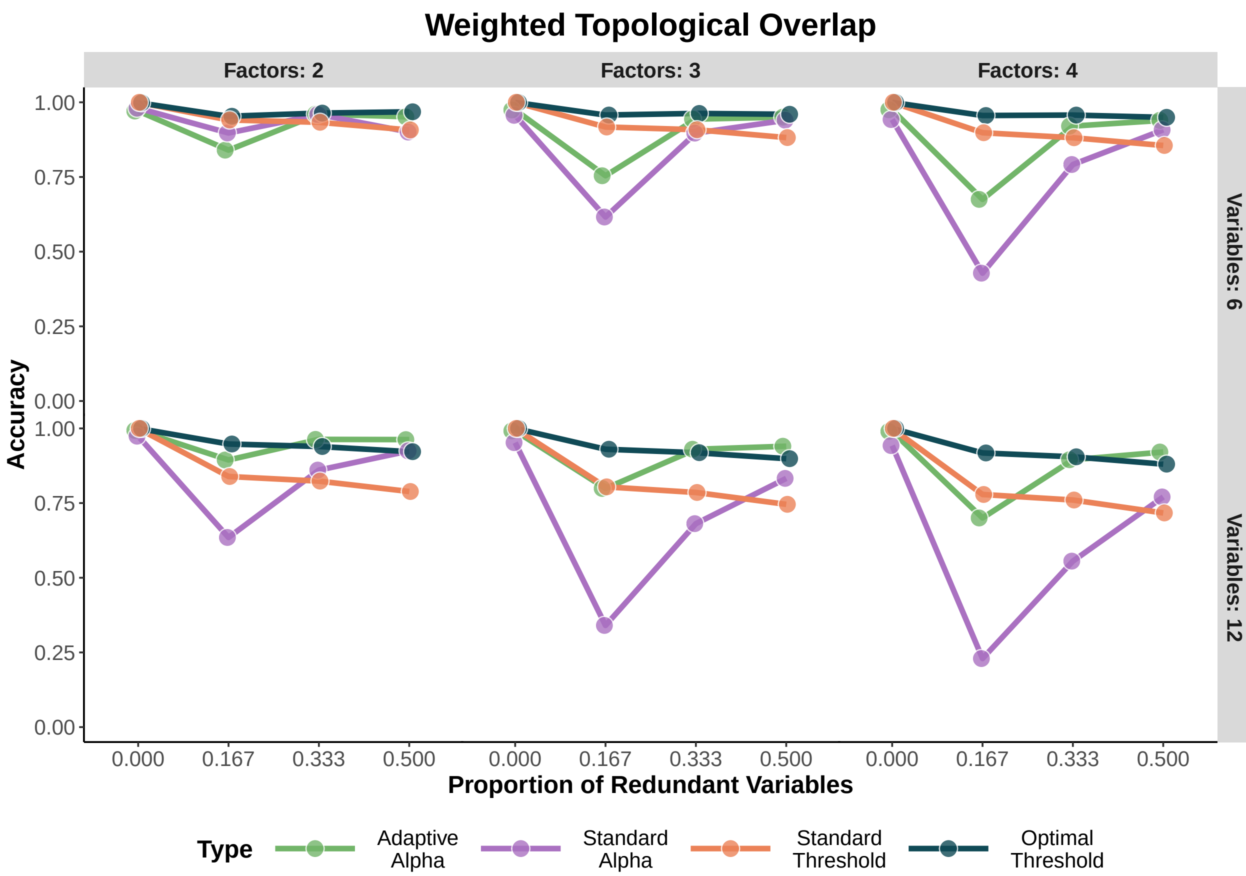

Figure 5

Accuracy of weighted topological overlap and threshold types. Factors = number of factors, Variables = number of variables per factor.

The interactions for both the number of factors and number of variables per factor were largely driven by when the proportion of redundant variables were small (0.167). The alpha thresholds in particular tended to decrease in accuracy as the number of factors increased and the number of variables per factor increased. This trend suggests that as the total number of variables increase (i.e., number of factors × number of variables per factor), the alpha thresholds' accuracy tends to decrease. The optimal threshold tended to perform the best when there were fewer total variables and lower redundant proportions (0.000 and 0.167), while the adaptive alpha threshold tended to perform the best when there more total variables and higher redundant proportions (0.333 and 0.500).

Overall

When examining the overall accuracy, there were no main effects or interactions. Therefore, we broken down accuracy by method, threshold type, and proportion of redundant variables (Figure 6).

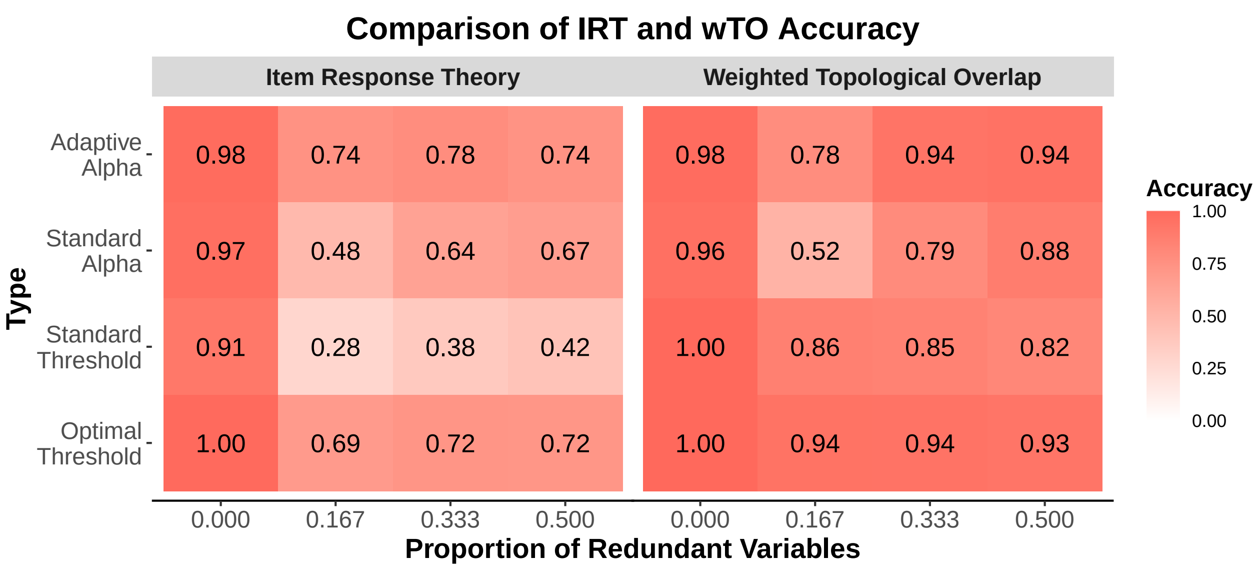

Figure 6

Overall accuracy of IRT and weighted topological overlap (wTO) threshold types.

The weighted topological overlap combinations with threshold types were all more accurate relative to their IRT counterpart (except for when there were no redundant variables). For IRT, the adaptive alpha threshold was the most accurate across all proportions of redundant values (except for when there were no redundant variables) including the optimal threshold. Despite evidence for the effectiveness of IRT's standard threshold in the literature, it had the worst accuracy for all conditions. This finding suggests that IRT's standard threshold of .20 is not viable for multidimensional structures. For weighted topological overlap, the optimal threshold was the most accurate across all proportion of redundant variables for all methods and threshold types (except 0.500). The standard threshold performed well when there were few redundant variables (0.000 and 0.167) but had the worst accuracy for weighted topological overlap when there were a lot of redundant variables (0.500).

Applied Example

For our applied example, we apply a new redundancy detection approach, UVA, to a dataset collected as a part of the Simons Foundation Autism Research Initiative's Simplex Collection (https://www.sfari.org/). This dataset included the *Broad Autism Phenotype Questionnaire* (BAPQ; Hurley, Losh, Parlier, Reznick, & Piven, 2007), which was completed by 5,659 people (fathers and mothers of a child with an autism spectrum disorder). The original dimensionality structure proposed by Hurley, Losh, Parlier, Reznick, and Piven (2007) has three factors which capture different aspects of the broad autism phenotype: aloof personality (represents a limited interest in or enjoyment of social interactions), rigid personality (refers to resistance, and/or difficulty adapting, to change) and pragmatic language (deficits in the social use of language leading to difficulties with effective communication and/or conversational reciprocity). Each factor has 12 items (36 items in total; for item descriptions, see SI8).

We begin by first estimating the dimension structure of the BAPQ inventory using EGA. We demonstrate that redundancy among some of the variables affects the recovery of the expected theoretical structure. After, we apply UVA. Our tutorial presents an automated UVA algorithm, which identifies redundant variables and merges them together as latent variables (following testlet response theory approach in IRT; Wainer, Bradlow, & Wang, 2007; Yen, 1984). UVA can also be applied manually, giving researchers the definitive decision to handle redundancy based on their theoretical knowledge about the relations between the variables.

Handling Redundancy

Following A. P. Christensen, Golino, and Silvia (2020), `UVA` provides two main approaches for reducing redundancy in data: removing all but one redundant variable and creating composite (sum scores) or latent variables from redundant variables. For the former approach, researchers select one variable from variables that are determined to be redundant and remove the other variables from the dataset. As a general heuristic, researchers can compute corrected item-test correlations for the variables in the redundant response set. The variable that has the largest correlation is likely to be the one that best captures the overall essence of the redundant variables (DeVellis, 2017; McDonald, 1999). Other rules of thumb for this approach are to select variables that have the most variance (DeVellis, 2017) and variables that are more general (e.g., "I often express my opinions" is better than "I often express my opinions in meetings" because it does not imply a specific context).

For the latter approach, redundant variables can be combined into a reflective latent variable and latent scores are estimated, replacing the redundant variables (Wainer, Bradlow, & Wang, 2007). Following recent suggestions for ordinal data with categories fewer than six, the Weighted Least Squares Mean- and Variance-adjusted (WLSMV) estimator is used; otherwise, if all categories are greater than or equal to six then Maximum Likelihood with Robust standard errors (MLR) is used (Rhemtulla, Brosseau-Liard, & Savalei, 2012). We strongly recommend the latent variable approach because it minimizes measurement error and retains all possible information available in the data.

Demonstration

To get started, the reader must install the latest *EGAnet* (Golino & Christensen, 2020) package from GitHub. After, the package can be loaded. Data for the BAPQ can be obtained from the Simons Foundation Autism Research Initiative's Simplex Collection (https://www.sfari.org/).

```r
# Download latest EGAnet package
devtools::install_github(
  "hfgolino/EGAnet",
  dependencies = c("Imports", "Suggests"),
  build_vignettes = TRUE
)
```

```r
# Load packages
library(EGAnet)
```

Initial Dimensionality Estimation

Moving forward with the application of the `UVA`, we start by evaluating the dimensional structure of the Broad Autism Phenotype Questionnaire (BAPQ) *without* reducing redundancy. The following code can be run:

```r
# EGA (with redundancy)
ega.wr <- EGA(
  data = bapq.all[,c(4:39)],
  plot.EGA = FALSE
)


plot.ega.wr <- plot(
  ega.wr,
```

```
  plot.args = list(node.size = 8, label.size = 4)
)


# Bootstrap EGA
boot.ega.bapq.all <- bootEGA(
  data = bapq.all[,c(4:39)],
  iter = 500, model = "glasso",
  ncores = 4
)


# Item Stability
itemStab.bapq <- itemStability(boot.ega.bapq.all)
```

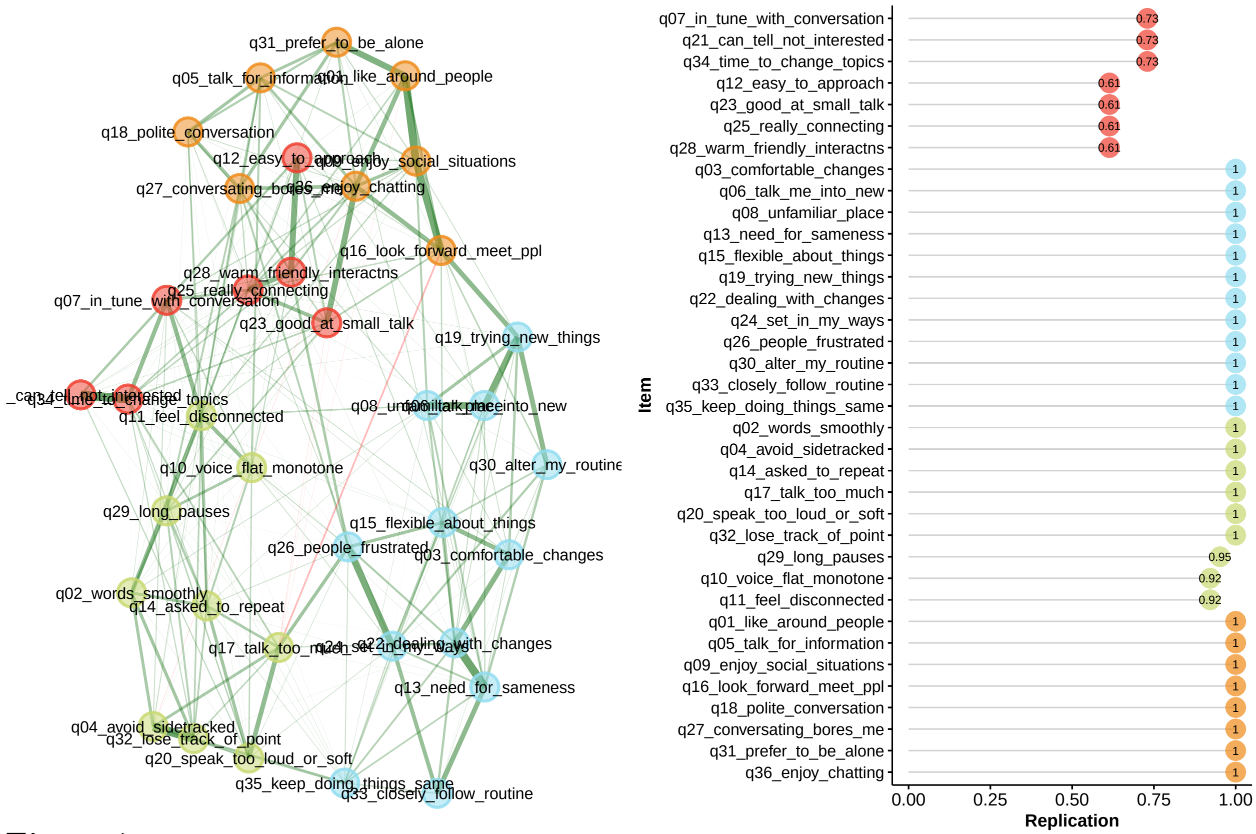

Figure 7

Exploratory Graph Analaysis of BAPQ questionnaire Before Unique Variable Analysis and item stability

Without performing UVA, EGA estimates that there are four factors (Figure 7). Investigating the items' descriptions of these factors (SI8), they represent roughly social interest (red; items 7, 12, 21, 23, 25, 28 and 34), rigid personality (blue; items 3, 6, 8, 13, 15, 19, 22, 24, 26, 30, 33 and 35), problems with pragmatic language (green; items 2, 4, 10, 11, 14, 17, 20, 29, 32), and aloof (orange; items 1, 5, 9, 16, 18, 27, 31, 36). These four factors are closely related to the original three-factor structure of the instrument (aloof personality, rigid personality, and problems with the pragmatic use of language; Hurley, Losh, Parlier, Reznick, & Piven, 2007). However, investigating the stability of the items to replicate in the original EGA structure, using bootstrap exploratory graph analysis implemented using a parametric bootstrap with 500 replica samples (A. P. Christensen & Golino, 2021), seven items present a replicability lower than 70%, which is below the suggested item replicability of 75% as suggested by A. P. Christensen and Golino (2021). Further, the structural consistency—the extent to which each dimension replicates with the exact same items—of this factor was low: 61.4% (A. P. Christensen, Golino, & Silvia, 2020). Overall, EGA suggested a four-factor structure for 73% of the bootstrap samples, and three factors in the remaining 27% of the cases.

UVA

To handle the redundancy in the scale, we can now use the `UVA` function:

```
# Perform automated UVA (latent variable)
bapq.auto <-  UVA(
  data = bapq.all[,c(4:39)],
  sig = .20, # Threshold value
  auto = TRUE
)
```

Running the code above will run through UVA's automated redundant detection and merging process. This process applies the weighted topological overlap with threshold to first identify variables that are redundant. Redundant variable sets are estimated as latent variables. The procedure then verifies that there are no leftover redundancies by applying the weighted topological overlap with threshold to the all remaining variables (empirical and latent). Any remaining redundancies are estimated as latent variables. If extant latent variables are redundant with other (latent) variables, then all variables that reflect the original latent variable(s) are used to re-estimate a new latent variable to replace the old one(s). This procedure continues iteratively until there are no redundant variables that remain. After the procedure completes, users have the opportunity to rename the latent variables to their desire.

Five item-pairs were identified as redundant. LV_1 (wTO = 0.25) was items 1 (*I like being around other people*) and 9 (*I enjoy being in social situations*), LV_2 (wTO = 0.22) was items 4 (*It's hard for me to avoid getting sidetracked in conversation*) and 32 (*I lose track of my original point when talking to people*), LV_3 (wTO = 0.22) was items 13 (*I feel a strong need for sameness from day to day*) and 22 (*I have a hard time dealing with changes in my routine*), LV_4 (wTO = 0.21) was items 21 (*I can tell when someone is not interested in what I am saying*) and 34 (*I can tell when it is time to change topics in conversation*), and LV_5 (wTO = 0.20) was items 6 (*People have to talk me into something new*) and 8 (*I have to warm myself up to the idea of visiting an unfamiliar place*; Figure 8).

Re-estimation of Dimensionality

UVA reduced the dataset from 36 items down to 31 items or sets of items (i.e., latent variables).

```
# EGA (with redundant variables combined)
ega.bapq.reduced <- EGA(bapq.auto$reduced$data)
plot.ega.reduced <- plot(
  ega.bapq.reduced,
  plot.args = list(node.size = 8, label.size = 4)
)



boot.ega.bapq.reduced <- bootEGA(
  data = bapq.auto$reduced$data,
  iter = 500, model = "glasso",
  ncores = 4
)


itemStab.bapq.reduced <- itemStability(boot.ega.bapq.reduced)
```

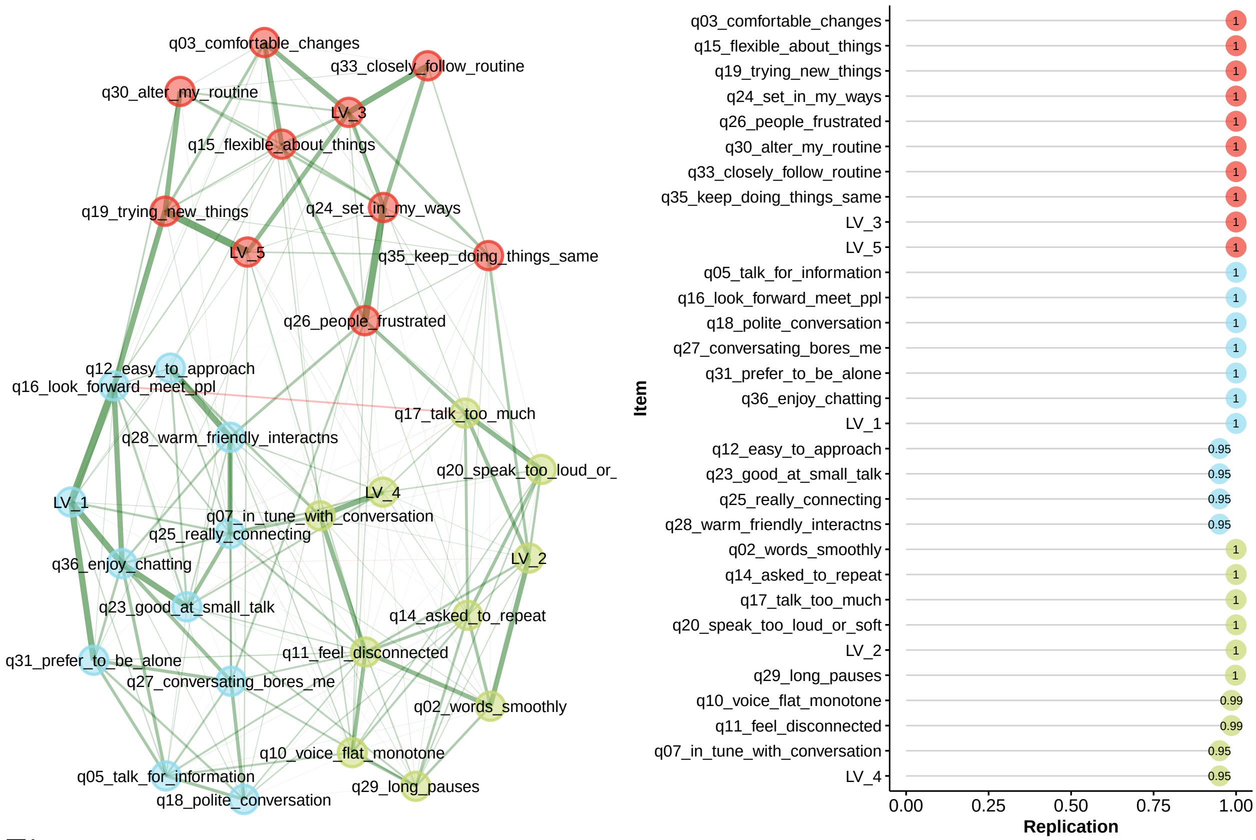

Figure 8

Exploratory Graph Analaysis of BAPQ questionnaire After Unique Variable Analysis and item stability

With this dataset, we then re-estimated the dimensionality of the BAPQ inventory using EGA. This time, three factors were estimated: rigid (red), aloof (blue), and pragmatic language (green; Figure 8). These three factors corresponded to the expected factor structure of the BAPQ questionnaire, corroborating the effectiveness of the UVA. In sum, our example demonstrates that redundancy can lead to instability of item stability, which may bias dimensionality estimates towards overfactoring (as shown in Figure 7). When this redundancy is handled, then the dimensionality estimates can be expected to be more accurate and in line with theoretical expectations (as shown in Figure 8), and the stability of the items is much higher, with a replicability above or equal to 95%. Finally, for this reduced item set EGA suggested a three-factor structure for the vast majority of the bootstrap samples (95%).

Discussion

The present paper is the first, to our knowledge, to develop an approach for detecting item redundancy that does not require the estimation of a latent structure or have the assumption of undimensionality. Our approach relied on a simple principle: use a threshold on the weighted topological overlap values of a psychometric network. Our simulation compared this approach to the more commonly applied Yen's Q3 in IRT. To avoid estimating dimensions, we used a unidimensional IRT model. We found that the weighted topological overlap approach, particularly with thresholds of .20 and .25, greatly outperformed the IRT approach. Importantly, we demonstrated that redundancy can have substantial effects on the estimation of the theoretical dimensional structure (major factors). Finally, we applied an automated algorithm of our approach, which we call Unique Variable Analysis, to real-world data. In our empirical example, we show the effects of redundancy on dimensionality estimates and how UVA can mitigate these effects to deliver dimensionality estimates that are more in line with theoretical expectations.

The results of the dimensionality analysis in our simulation revealed strong effects of redundancy. In many regards, this result was not surprising. Our data generating method purposely manipulated a certain proportion of these variables to be redundant. The manipulated variables were adjusted to mimic response patterns of clear and obvious redundancy in real-world data. Importantly, we did not manipulate other variables, making the difference between redundant and non-redundant variables clearer than would be expected in real-world data. Our approach thus changed the known factor structure and perhaps exaggerated the magnitude of the redundancy effects on dimensionality relative to real-world data (especially with larger proportions of redundancy). Nonetheless, our goal was to demonstrate the potential problems redundancy can have on the estimation of expected dimension structures.

Despite the potential limitations of the smaller gradation between non-redundant and redundant variables, our data generation approach was optimal for evaluating the weighted topological overlap approach and a local dependence IRT statistic. Based on previous research, thresholds for both weighted topological overlap (.30; Nowick, Gernat, Almaas, & Stubbs, 2009) and IRT (.20; Yen, 1984) were already present in the literature. Using a grid search, we explored whether these thresholds were the best for our generated data. We found that .20 and .25 for weighted topological overlap and .35 for IRT had higher accuracies than these previously reported standards. For IRT, it's important to emphasize that we estimated a unidimensional model on multidimensional data. This application is in violation of the fundamental unidimensional assumption of the model. Our results provide clear evidence of this violation: as correlations between factors increased, the accuracy of the IRT approach also increased. Correlations between factors can be used as a rough proxy for the distance of a multidimensional structure from a unidimensional structure. The larger the correlations between factors, the more a multidimensional structure will be "like" a unidimensional structure. To this end, we cannot rule out the commonly applied threshold of .20 for local dependence in unidimensional structures.

For overall redundancy detection performance, the weighted topological overlap outperformed IRT. We examined several threshold strategies that included conventional thresholding and significance-based (percentile) thresholding. Across all thresholds, weighted topological overlap outperformed IRT. The optimal threshold for weighted topological overlap had the best accuracy across threshold types for the approach. For IRT, the adaptive alpha threshold had the highest accuracy, outperforming the optimal threshold identified by our grid search. This result suggests that, at least for multidimensional structures, the adaptive alpha threshold might be preferred over the optimal and standard threshold.

Based on our simulation results, we developed an automated algorithm to detect and merge redundant variables. We then used the automated UVA procedure on an empirical dataset, the BAPQ, demonstrating how redundancy or in this case the underlying semantic similarity between items on different theoretical factors affects its dimension stability. After the application of UVA, the theoretical, expected structure is returned by empirical EGA and it is demonstrated to have a robust dimension structure (i.e., stable dimensions). In our example, we used the threshold .20, which differs from what was deemed optimal by our grid search. We note, however, that our grid search demonstrated that thresholds of .20 and .25 were highly comparable and only differed on the lowest (0.167) and highest (0.500) redundancy proportions (see Figure 3). Our choice to use the threshold of .20 rather than .25 was based on the results of the dimension stability analysis. When using the threshold of .25, only one pair of variables was redundant and the dimensional structure remained unstable (there was factor splitting). Using the threshold of .20 handled this issue and produced results that were consistent with the theoretically intended structure. In practice, we propose that researchers apply both thresholds to examine the results. Researchers should then use theoretical knowledge to guide their decision for their results.

One general takeaway researchers might have from our presentation, simulation, and empirical representation of redundancy is that all redundancy is bad. This does *not* represent our view. Our view is that redundancy can be represented on a spectrum. The simulation employed in this study examined particularly strong redundancy between variables—so much so that dimensionality estimates were affected. Redundancy to a lesser degree may not affect dimensionality; however, it may affect parameter estimates. Local dependency, even dependencies weaker than what we examined, can have effects on item parameter estimates in IRT (W.-H. Chen & Thissen, 1997). Similarly, network parameters are likely to be affected. For networks, redundancy can contribute to the extent to which two variables connected to each other and other variables, which consequently affects network measures such as node strength (or the sum of a node's connections). Redundancy can therefore make a node appear particularly well connected in a network when the connections are due to similarities between item wordings rather than actual substantive importance. Future work is necessary to explore the redundancy spectrum to identify when redundancy becomes a problem at different levels (e.g., item parameters, dimensionality, hierarchical structures; W.-H. Chen & Thissen, 1997; Hallquist, Wright, & Molenaar, 2019). Different thresholds are likely to be effective for different redundancy problems. Some researchers may be interested in eliminating redundancy effects on item parameters whereas others may only be concerned with its effects on dimensionality estimates.

Dimensionality, as we show, can be strongly affected by redundancy which may lead to lower consistency of the dimensions across studies, leading to problems in understanding the substantive interpretation of the dimensions (Gerbing & Anderson, 1984; Goldstein, 1980). One reason for unreliable dimensions may be that minor factors or split factors are appearing in some studies but not others. Some of the more consequential effects may appear when attempting to measure broad attributes such as personality traits where some facets contain substantial redundancy while others do not (A. P. Christensen, Golino, & Silvia, 2020). Oftentimes, these issues appear because of shared semantic similarity where some items are slight variations in wording and content (Leising et al., 2020; Rosenbusch, Wanders, & Pit, 2020). Undoubtedly, semantic similarity between items will create the appearance of more internally consistent scales but at the cost of narrowing the scale content, and thus changing the nature and scope of the construct being assessed (Clark & Watson, 2019).

After detection, item redundancy can be handled in different ways depending on the statistical analyses being performed. When conducting a dimensionality assessment, item redundancy can be addressed by creating latent variables for the redundant variables, or by eliminating all but one variable of the redundant item groups. For IRT models, redundant variables can used to create testlets that are unidimensional (Wainer, Bradlow, & Wang, 2007). For psychometric network models, redundancy requires specific attention because parameter estimates are likely to be skewed (A. P. Christensen, Golino, & Silvia, 2020; Hallquist, Wright, & Molenaar, 2019). In this case, item redundancy can be handled in a similar way to dimensionality assessment. Our Unique Variable Analysis is designed to inform researchers about the potential redundancies in their data and is equipped with an automated procedure to handle those redundancies.

Researchers should endeavor to reduce item redundancy as much as possible in psychometric testing. As an example, consider the *frustration leads to aggression* hypothesis (Nesselroade & Molenaar, 2016): Students may became frustrated when trying to solve difficult mathematical problems; some may curse the teacher while others will walk away from the class, and still some students might slap a fellow student. Cursing, walking away, and slapping someone are, in a causal latent variable perspective, three indicators of a general behavioral tendency that we can label as *"frustration produces aggression"*. The three behaviors are distinctively unique, and good indicators of the latent tendency to became aggressive due to frustration. Their correlation may be moderate-to-high, and together these indicators might produce a factor with high internal consistency while avoiding item redundancy.

This example shows that it is possible to estimate latent factors providing the generality needed to articulate useful lawfulness without having to rely on redundant indicators (i.e., without "cheating by repeating"; Reise, Bonifay, & Haviland, 2018). The need to use redundant indicators, as in a shared semantic reference approach (i.e., using similar item phrasing or similar item content; Rosenbusch, Wanders, & Pit, 2020), may be a well-known scale construction strategy, but it reflects either a lack of strong theory to guide the scale construction or the development of scales with artificially inflated internal consistency (Reise, Bonifay, & Haviland, 2018). For example, *"I like going to parties"*, *"I like attending social events"* and *"I'm very fond of gatherings and festivities"* may reflect extraversion, but won't provide much differential information regarding this personality trait, since the three items are basically identical in terms of content and simply use different synonyms for *parties*. Our approach can be used to help the identification of redundant items with an eye towards the identification of generalizable latent constructs.

When developing tests, redundancy assessment, such as the novel approach developed here, should be integrated into other test validation practices (Flake, Pek, & Hehman, 2017). We view redundancy as the first step of structural validation (A. P. Christensen, Golino, & Silvia, 2020). Importantly, our approach allows redundancy to be determined prior to estimating factors, which mitigates confounding effects of different factor structures. Our simulation demonstrated that our novel approach is effective for detecting redundancy without knowledge of the latent structure of the data (unlike traditional local dependence measures) and our empirical example demonstrated how researchers can apply our approach to their data. Across our paper, we've discussed and demonstrated that redundancy can have substantial unintended effects if left unchecked. Future work is necessary to understand what other consequences might exist for redundancy in data (e.g., item parameter estimation, measurement precision; Markon, 2013). Our hope is that researchers will become more aware of these effects and their potential consequences.

References

Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown.* Retrieved from https://github.com/crsh/papaja

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*, P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health*, *6*, 149. https://doi.org/10.3389/fpubh.2018.00149

Bollmann, S., Heene, M., Küchenhoff, H., & Bühner, M. (2015). *What can the real world do for simulation studies? A comparison of exploratory methods.* Retrieved from Department of Statistics, University of Munich: Retrieved from https://epub.ub.uni-muenchen.de/24518/

Boyle, G. J. (1991). Does item homogeneity indicate internal consistency or item redundancy in psychometric scales? *Personality and Individual Differences*, *12*, 291–294. https://doi.org/10.1016/0191-8869(91)90115-R

Cattell, R. B. (1978). *The scientific use of factor analysis in behavioral and life sciences.* Boston, MA: Springer. https://doi.org/10.1007/978-1-4684-2262-7

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Champely, S. (2020). *pwr: Basic functions for power analysis.* Retrieved from https://CRAN.R-project.org/package=pwr

Chen, J., & Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, *95*, 759–771. https://doi.org/10.1093/biomet/asn034

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289. https://doi.org/10.3102/10769986022003265

Christensen, A. P., Garrido, L. E., & Golino, H. (2021). Comparing community detection algorithms in psychological data: A Monte Carlo simulation. *PsyArXiv*. https://doi.org/10.31234/osf.io/hz89e

Christensen, A. P., & Golino, H. (2021). Estimating the stability of psychological dimensions via Bootstrap Exploratory Graph Analysis: A Monte Carlo simulation and tutorial. *PsyArXiv*. https://doi.org/10.31234/osf.io/9deay

Christensen, A. P., Golino, H., & Silvia, P. J. (2020). A psychometric network perspective on the validity and validation of personality trait questionnaires. *European Journal of Personality*, *34*, 1095–1108. https://doi.org/10.1002/per.2265

Christensen, A. P., Gross, G. M., Golino, H., Silvia, P. J., & Kwapil, T. R. (2019). Exploratory Graph Analysis of the Multidimensional Schizotypy Scales. *Schizophrenia Research*, *206*, 43–51. https://doi.org/10.1016/j.schres.2018.12.018

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement*, *41*(3), 178–194. https://doi.org/10.1177/0146621616677520

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, *31*(12), 1412–1427. https://doi.org/http://dx.doi.org/10.1037/pas0000626

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Comrey, A. L. (1988). Factor-analytic methods of scale development in personality and clinical psychology. *Journal of Consulting and Clinical Psychology*, *56*, 754–761. https://doi.org/10.1037/0022-006X.56.5.754

Condon, D. M. (2018). The SAPA personality inventory: An empirically-derived, hierarchically-organized self-report personality assessment model. *PsyArXiv*.

https://doi.org/10.31234/osf.io/sc4p9

Cooper, C. (2019). Pitfalls of personality theory. *Personality and Individual Differences*, *151*, 109551. https://doi.org/10.1016/j.paid.2019.109551

Cramer, A. O. J. (2012). Why the item "23+ 1" is not in a depression questionnaire: Validity from a network perspective. *Measurement: Interdisciplinary Research & Perspective*, *10*, 50–54. https://doi.org/10.1080/15366367.2012.681973

Cramer, A. O. J., van der Sluis, S., Noordhof, A., Wichers, M., Geschwind, N., Aggen, S. H., ... Borsboom, D. (2012). Dimensions of normal personality as networks in search of equilibrium: You can't like parties if you don't like people. *European Journal of Personality*, *26*, 414–431. https://doi.org/10.1002/per.1866

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. https://doi.org/10.1007/BF02310555

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, *1695*, 1–9. Retrieved from https://www.semanticscholar.org/paper/The-igraph-software-package-for-complex-network-Cs/%C3/%A1rdi-Nepusz/1d2744b83519657f5f2610698a8ddd177ced4f5c?p2df

DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). Thousand Oaks, CA: SAGE Publications.

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, *16*(1), 5–18. https://doi.org/https://doi.org/10.1007/s11136-007-9198-0

Edwards, M. C., Houts, C. R., & Cai, L. (2018). A diagnostic procedure to detect departures from local independence in item response theory models. *Psychological Methods*, *23*(1), 138–149. https://doi.org/http://dx.doi.org/10.1037/met0000121

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D.

(2012). qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software*, *48*, 1–18. https://doi.org/10.18637/jss.v048.i04

Epskamp, S., & Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological Methods*, *23*, 617–634. https://doi.org/10.1037/met0000167

Epskamp, S., Maris, G., Waldrop, L. J., & Borsboom, D. (2018). Network psychometrics. In P. Irwing, D. Hughes, & T. Booth (Eds.), *The Wiley handbook of psychometric testing, 2 volume set: A multidisciplinary reference on survey, scale and test development*. New York, NY: Wiley. https://doi.org/10.1002/9781118489772.ch30

Fava, J. L., & Velicer, W. F. (1992). The effects of overextraction on factor and component analysis. *Multivariate Behavioral Research*, *27*(3), 387–415. https://doi.org/10.1207/s15327906mbr2703_5

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*, 370–378. https://doi.org/10.1177/1948550617693063

Flores-Kanter, P. E., Garrido, L. E., Moretti, L. S., & Medrano, L. A. (2021). A modern network approach to revisiting the positive and negative affective schedule (PANAS) construct validity. *Journal of Clinical Psychology*, 1–35. https://doi.org/https://doi.org/10.1002/jclp.23191

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*, 75–174. https://doi.org/10.1016/j.physrep.2009.11.002

Foygel, R., & Drton, M. (2010). Extended Bayesian information criteria for Gaussian graphical models. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, & A. Culotta (Eds.), *Advances in neural information processing systems* (pp. 604–612). Retrieved from https://papers.nips.cc/paper/4087-extended-bayesian-information-criteria-for-gaussian-graphical-models

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*, 432–441.

https://doi.org/10.1093/biostatistics/kxm045

Friedman, J., Hastie, T., & Tibshirani, R. (2014). *glasso: Graphical lasso – estimation of Gaussian graphical models*. Retrieved from https://CRAN.R-project.org/package=glasso

García-Garzón, E., Abad, F. J., & Garrido, L. E. (2019). Improving bi-factor exploratory modeling: Empirical target rotation based on loading differences. *Methodology*, *15*, 45–55. https://doi.org/10.1027/1614-2241/a000163

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2013). A new look at Horn's parallel analysis with ordinal variables. *Psychological Methods*, *18*, 454–474. https://doi.org/10.1037/a0030005

Garrido, L. E., Barrada, J. R., Aguasvivas, J. A., Martínez-Molina, A., Arias, V. B., Golino, H. F., . . . Rojo-Moreno, L. (2020). Is small still beautiful for the Strengths and Difficulties Questionnaire? Novel findings using exploratory structural equation modeling. *Assessment*, *27*, 1349–1367. https://doi.org/10.1177/1073191118780461

Gates, K. M., Henry, T., Steinley, D., & Fair, D. A. (2016). A Monte Carlo evaluation of weighted community detection algorithms. *Frontiers in Neuroinformatics*, *10*, 45. https://doi.org/10.3389/fninf.2016.00045

Gerbing, D. W., & Anderson, J. C. (1984). On the meaning of within-factor correlated measurement errors. *Journal of Consumer Research*, *11*, 572–580. https://doi.org/10.1086/208993

Glas, C. A., & Falcón, J. C. S. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, *27*(2), 87–106. https://doi.org/10.1177/0146621602250530

Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology*, *33*, 234–246. https://doi.org/10.1111/j.2044-8317.1980.tb00610.x

Golino, H., & Christensen, A. P. (2020). *EGAnet: Exploratory Graph Analysis– a framework for estimating the number of dimensions in multivariate data using network* *psychometrics*. Retrieved from https://CRAN.R-project.org/package=EGAnet

Golino, H., & Epskamp, S. (2017). Exploratory Graph Analysis: A new approach for estimating the number of dimensions in psychological research. *PLoS ONE*, *12*, e0174035. https://doi.org/10.1371/journal.pone.0174035

Golino, H., Shi, D., Christensen, A. P., Garrido, L. E., Nieto, M. D., Sadana, R., ... Martinez-Molina, A. (2020). Investigating the performance of Exploratory Graph Analysis and traditional techniques to identify the number of latent factors: A simulation and tutorial. *Psychological Methods*, *25*, 292–320. https://doi.org/10.1037/met0000255

Gysi, D. M., Voigt, A., de Miranda Fragoso, T., Almaas, E., & Nowick, K. (2018). wTO: An R package for computing weighted topological overlap and a consensus network with integrated visualization tool. *BMC Bioinformatics*, *19*, 392. https://doi.org/10.1186/s12859-018-2351-7

Hallquist, M., Wright, A. C. G., & Molenaar, P. C. M. (2019). Problems with centrality measures in psychopathology symptom networks: Why network psychometrics cannot escape psychometric theory. *Multivariate Behavioral Research*. https://doi.org/10.1080/00273171.2019.1640103

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, *7*(2), 191–205. https://doi.org/10.1177/1094428104263675

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement*, *66*, 393–416. https://doi.org/10.1177/0013164405282485

Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review*, *14*, 332–346. https://doi.org/10.1177/1088868310361240

Hurley, R. S. E., Losh, M., Parlier, M., Reznick, J. S., & Piven, J. (2007). The broad autism phenotype questionnaire. *Journal of Autism and Developmental Disorders*, *37*, 1679–1690.

https://doi.org/10.1007/s10803-006-0299-3

Lauritzen, S. L. (1996). *Graphical models*. Oxford, UK: Clarendon Press.

Leising, D., Burger, J., Zimmermann, J., Bäckström, M., Oltmanns, J., & Connelly, B. (2020). Why do items correlate with one another? A conceptual analysis with relevance for general factors and network models. *PsyArXiv*. https://doi.org/10.31234/osf.io/7c895

MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques. *MIS Quarterly*, *35*, 293–334. https://doi.org/10.2307/23044045

Marais, I. (2012). Local dependence. In K. B. Christensen, S. Kriner, & M. Mesbah (Eds.), *Rasch models in health* (pp. 111–130). London, UK: Wiley.

Markon, K. E. (2013). Information utility: Quantifying the total psychometric information provided by a measure. *Psychological Methods*, *18*, 15. https://doi.org/10.1037/a0030638

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, *52*, 81–90. https://doi.org/10.1037/0022-3514.52.1.81

McCrae, R. R., & Mõttus, R. (2019). What personality scales measure: A new psychometrics and its implications for theory and assessment. *Current Directions in Psychological Science*. https://doi.org/10.1177/0963721419849559

McDonald, R. P. (1999). *Test theory: A unified treatment*. New York, NY: Taylor & Francis. https://doi.org/10.4324/9781410601087

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*, 412–433. https://doi.org/10.1037/met0000144

Montoya, A. K., & Edwards, M. C. (2020). The poor fit of model fit for selecting number of factors in exploratory factor analysis for scale evaluation. *Educational and Psychological Measurement*. https://doi.org/10.1177/0013164420942899

Nesselroade, J. R., & Molenaar, P. C. M. (2016). Some behavioral science measurement concerns and proposals. *Multivariate Behavioral Research*, *51*, 396–412.

https://doi.org/10.1080/00273171.2015.1050481

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, *103*, 8577–8582. https://doi.org/10.1073/pnas.0601602103

Nowick, K., Gernat, T., Almaas, E., & Stubbs, L. (2009). Differences in human and chimpanzee gene expression patterns define an evolving network of transcription factors in brain. *Proceedings of the National Academy of Sciences*, *106*, 22358–22363. https://doi.org/10.1073/pnas.0911376106

Perezgonzalez, J. D. (2015). P-values as percentiles. Commentary on: "Null hypothesis significance tests. A mix–up of two different theories: The basis for widespread confusion and numerous misinterpretations." *Frontiers in Psychology*, *6*, 341. https://doi.org/10.3389/fpsyg.2015.00341

Pérez, M.-E., & Pericchi, L. R. (2014). Changing statistical significance with the amount of information: The adaptive α significance level. *Statistics & Probability Letters*, *85*, 20–24. https://doi.org/10.1016/j.spl.2013.10.018

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., & Barabási, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, *297*, 1551–1555. https://doi.org/10.1126/science.1073374

Reise, S. P., Bonifay, W., & Haviland, M. G. (2018). Bifactor modelling and the evaluation of scale scores. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing* (pp. 675–707). Chichester, UK: Wiley. https://doi.org/10.1002/9781118489772.ch22

Revelle, W. (2019). *psychTools: Tools to accompany the 'psych' package for psychological research*. Evanston, Illinois: Northwestern University. Retrieved from https://CRAN.R-project.org/package=psychTools

Revelle, W., Dworak, E. M., & Condon, D. M. (2020). Exploring the persome: The power of the item in understanding personality structure. *Personality and Individual Differences*, 109905. https://doi.org/10.1016/j.paid.2020.109905

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*, 354–373. https://doi.org/10.1037/a0029315

Rosenbusch, H., Wanders, F., & Pit, I. L. (2020). The Semantic Scale Network: An online tool to detect semantic overlap of psychological scales and prevent scale redundancies. *Psychological Methods*, *25*, 380–392. https://doi.org/10.1037/met0000244

Saris, W. E., Satorra, A., & Veld, W. M. van der. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling: A Multidisciplinary Journal*, *16*, 561–582. https://doi.org/https://doi.org/10.1080/10705510903203433

Saucier, G., & Iurino, K. (2020). High-dimensionality personality structure in the natural language: Further analyses of classic sets of english-language trait-adjectives. *Journal of Personality and Social Psychology*, *119*, 1188–1219. https://doi.org/10.1037/pspp0000273

Schloerke, B., Cook, D., Larmarange, J., Briatte, F., Marbach, M., Thoen, E., … Crowley, J. (2020). *GGally: Extension to 'ggplot2'*. Retrieved from https://CRAN.R-project.org/package=GGally

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge, UK: Cambridge University Press.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and*

Social Psychology, *54*, 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York, NY: Springer. Retrieved from https://ggplot2-book.org/

Wood, J. M., Tataryn, D. J., & Gorsuch, R. L. (1996). Effects of under- and overextraction on principal axis factor analysis with varimax rotation. *Psychological Methods*, *1*(4), 354–365. https://doi.org/10.1037/1082-989X.1.4.354

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*, 125–145. https://doi.org/10.1177/014662168400800201

Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, *4*, 17. https://doi.org/10.2202/1544-6115.1128

Supplementary Information

Unique Variable Analysis: A Novel Approach for Detecting Redundant Variables in Multivariate Data

SI 1. SAPA Inventory

Our approach to simulate redundancy started with descriptive analyses of known redundancy in a large real-world dataset. We used the Synthetic Aperture Personality Assessment (SAPA) inventory (Condon, 2018) dataset available in the *psychTools* (Revelle, 2019) package in R. A subset of these items (n = 70) form a five-factor structure that corresponds to the Five Factor Model (FFM; McCrae & Costa, 1987). The response options ranged from 1 (very inaccurate) to 6 (very accurate).

This 70-item subset was completed by 4,000 participants over the SAPA project website (https://sapa-project.org). These participants were collected after the developmental dataset (from February 2017 to May 2017) and were the first 4000 complete cases (rather than the first 4000 participants; D. Condon, personal communication, January 29, 2020). The sample had a mean age of 26.90 (SD = 11.49, range = 11–90) and were well represented for both sex (59.5% female) and education (11.1% graduated high school, 31.8% currently in university, 22% graduated university, and 11.8% held a graduate or professional degree). Race and ethnicity demographics were not provided; however, the data was gathered via the SAPA project website allowing equal opportunity for people of all ages, genders, ethnicities, and socio-economic backgrounds as long as they had access to the internet.

One potential sampling bias for this sample was that these participants were included because they completed all 135 items meaning that participants who did not complete all 135 items during the same time period were not included (regardless of whether they stopped or unintentionally skipped an item; D. Condon, personal communication, January 29, 2020). Despite this potential for representative bias, this sample likely represents a broader and more diverse population than most other self-report research in the personality literature.

There were several reasons for choosing this dataset, but we elaborate on a few only. First, the dataset is a large, diverse sample that is open-source, making the analyses performed in this study available for replication. Second, personality and self-report inventories are perhaps the most commonly used assessment instruments across psychological research and therefore represent the vast majority of the applications that the UVA will be used for. Finally, the SAPA inventory is structured hierarchically: there are 27 empirically derived lower-order dimensions that can be further collapsed into the prototypical FFM (Condon, 2018). These lower-order dimensions contain substantial redundancy, making the dataset a good example for how UVA can be applied and the number of unique components to expect after reducing the redundancy of the inventory (i.e., around 27). In addition, although the true dimensional structure cannot be known, there is strong theoretical support that five dimensions should underlie the data.

SI 2. SAPA Item Label and Description

Item Label	Description
q_90	Am concerned about others.
q_1763	Sympathize with others feelings.
q_253	Am sensitive to the needs of others.
q_1896	Use others for my own ends.
q_851	Feel sympathy for those who are worse off than myself.
q_1832	Think of others first.
q_501	Cheat to get ahead.
q_377	Believe that others have good intentions.
q_871	Feel that most people cant be trusted.
q_1855	Trust what people say.
q_4296	Tell a lot of lies.
q_142	Am hard to satisfy.
q_379	Believe that people are basically moral.
q_4289	Trust people to mainly tell the truth.
q_1290	Like order.
q_1744	Start tasks right away.
q_1979	Work hard.
q_1452	Neglect my duties.
q_1915	Want every detail taken care of.
q_1201	Keep things tidy.
q_530	Continue until everything is perfect.
q_904	Find it difficult to get down to work.
q_1867	Try to follow the rules.
q_1694	Set high standards for myself and others.
q_369	Believe that laws should be strictly enforced.

Item Label	Description
q__1444	Need a push to get started.
q__1483	Often forget to put things back in their proper place.
q__1254	Leave a mess in my room.
q__979	Get overwhelmed by emotions.
q__4252	Am a worrier.
q__1989	Worry about things.
q__1505	Panic easily.
q__4249	Would call myself a nervous person.
q__808	Fear for the worst.
q__793	Experience my emotions intensely.
q__1840	Think that my moods dont change more than most peoples do.
q__811	Feel a sense of worthlessness or hopelessness.
q__1585	Rarely get irritated.
q__578	Dislike myself.
q__176	Am not easily annoyed.
q__797	Experience very few emotional highs and lows.
q__1683	Seldom get mad.
q__1904	Usually like to spend my free time with people.
q__4243	Like going out a lot.
q__312	Avoid company.
q__565	Dislike being the center of attention.
q__1416	Make myself the center of attention.
q__1923	Want to be left alone.
q__1027	Hate being the center of attention.
q__684	Dont like crowded events.

Item Label	Description
q_254	Am skilled in handling social situations.
q_1296	Like to attract attention.
q_901	Find it difficult to approach others.
q_1243	Laugh a lot.
q_803	Express myself easily.
q_1244	Laugh aloud.
q_128	Am full of ideas.
q_2745	Am able to come up with new and different ideas.
q_2754	Am an original thinker.
q_1392	Love to think up new ways of doing things.
q_1058	Have a vivid imagination.
q_240	Am quick to understand things.
q_1738	Spend time reflecting on things.
q_422	Can handle a lot of information.
q_1389	Love to reflect on things.
q_1310	Like to get lost in thought.
q_1880	Try to understand myself.
q_747	Enjoy being thought of as a normal mainstream person.
q_1609	Rebel against authority.
q_1834	Think quickly.

SI 3. Known Redundant Items in the SAPA Inventory

(R) = reverse of parent label

Attention-seeking

"Dislike being the center of attention" (R), "Hate being the center of attention" (R), "Make myself the center of attention," and "Like to attract attention"

Orderly

"Keep things tidy," "Leave a messing my room" (R), "Like order," and "Often forget to put things back in their place" (R)

Worrier

"Am a worrier" and "Worry about things"

Laughter

"Laugh a lot" and "Laugh aloud"

Others-oriented

"Sympathize with others' feelings," "Am sensitive to the needs of others," "Am concerned about others," and "Think of others first"

Trusting

"Trust people to mainly tell the truth," "Trust what people say," and "Feel that most people can't be trusted" (R)

Sees good in people

"Believe that others have good intentions" and "Believe that people are basically moral"

Novel thinker

"Am an original thinker" and "Love to think up new ways of doing things"

Motivated

"Start tasks right away" and "Find it difficult to get down to work" (R)

High emotionality

"Experience my emotions intensely" and "Experience very few emotional highs and lows" (R)

Anxious

"Panic easily" and "Fear for the worst"

Even-tempered

"Rarely get irritated," "Seldom get mad," and "Am not easily annoyed"

People person

"Like going out a lot," "Avoid company" (R), "Usually like to spend my free time with people," and "Want to be left alone"

Socially skilled

"Am skilled in handling social situations" and "Find it difficult to approach others" (R)

Unconventional

"Enjoy being thought of as a normal mainstream person" (R) and "Rebel against authority"

SI 4. GLASSO Estimation and Exploratory Graph Analysis

Exploratory graph analysis (EGA) is a recently developed method to estimate the number of dimensions in multivariate data using undirected network models (Golino & Epskamp, 2017; Golino et al., 2020). EGA first applies a network estimation method followed by a community detection algorithm for weighted networks (Fortunato, 2010). EGA has been shown to be as accurate or more accurate than more traditional factor analytic methods such as parallel analysis (A. P. Christensen, Garrido, & Golino, 2021; Golino et al., 2020). EGA was applied using the *EGAnet* package (version 0.9.9; Golino & Christensen, 2020) in R (R Core Team, 2020).

Network Estimation Method

This study applied the graphical least absolute shrinkage and selection operator (GLASSO; Friedman, Hastie, & Tibshirani, 2008; Friedman, Hastie, & Tibshirani, 2014), which estimates a Gaussian Graphical Model (GGM; Lauritzen, 1996) where nodes (circles) represent variables and edges (lines) represent the conditional dependence (or partial correlations) between nodes given all other nodes in the network. The least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996) of the GLASSO is a regularization technique that reduces parameter estimates with some estimates becoming exactly zero.

The LASSO uses a parameter called lambda (λ), which controls the sparsity of the network. Lower values of λ remove fewer edges, increasing the possibility of including spurious correlations, while larger values of λ remove more edges, increasing the possibility of removing relevant edges. When $\lambda = 0$, then the estimates are equal to the ordinary least squares solution for the partial correlation matrix. In this study, the ratio of the minimum and maximum λ was set to 0.1.

The popular approach in the network psychometrics literature is to compute models across several values of λ (usually 100) and to select the model that minimizes the extended

Bayesian information criterion (EBIC; J. Chen & Chen, 2008; Epskamp & Fried, 2018). In network psychometrics literature, this approach has been termed EBICglasso and is applied using the *qgraph* package (Epskamp, Cramer, Waldorp, Schmittmann, & Borsboom, 2012) in R. The EBIC model selection uses a hyperparameter gamma (γ) to control how much it prefers simpler models (i.e., models with fewer edges; Foygel & Drton, 2010). Larger γ values lead to simpler models, while smaller γ values lead to denser models. When $\gamma = 0$, the EBIC is equal to the Bayesian information criterion. Following the EGA approach (Golino et al., 2020), if there was a disconnected node in the network, then γ was decreased by 0.25 until all nodes had at least one connection or γ reached zero.

Community Detection Algorithm

The Louvain algorithm (also referred to as Multi-level; Blondel, Guillaume, Lambiotte, & Lefebvre, 2008) is one of the most commonly applied community detection algorithms in network science (Gates, Henry, Steinley, & Fair, 2016). The algorithm begins by randomly sorting nodes into communities with their neighbors and then uses modularity (Newman, 2006) to iteratively optimize its community partitions by exchanging nodes between communities and evaluating the change in modularity until it no longer improves. Then, the algorithm collapses the communities into latent nodes and identifies edge weights with other observed and latent nodes, which provides a multi-level structure (Gates, Henry, Steinley, & Fair, 2016). In this study, the algorithm was not used to identify hierarchical community structures in the network. The Louvain algorithm was implemented using the *igraph* package (Csardi & Nepusz, 2006) in R. It's also important to note that the algorithm implemented in igraph is deterministic; however, other implementations are not (Gates, Henry, Steinley, & Fair, 2016).

SI 5. Adaptive Alpha

The standard alpha simply selects all weights that have a p-value less than .05. Adaptive alpha adjusts the standard alpha level by accounting for a reference sample size. It's well-known that as sample size increases, the likelihood of a small effect becoming significant also increases. To account for this, Pérez and Pericchi (2014) provide the following formula:

$$\alpha_{adapt} = \frac{\alpha * \sqrt{n_0 \times (\log(n_0) + \chi^2_\alpha(1))}}{\sqrt{n^* \times (\log(n^*) + \chi^2_\alpha(1))}},$$

where n_0 is the reference sample size, n^* is the actual sample size, and α is the standard alpha (i.e., $\alpha = .05$). The reference sample size can be computed using a power analysis. For our purposes, this power analysis was computed using the *pwr* package (Champely, 2020) in R for a correlation with a medium effect size ($r = .30$), alpha of .05, and power of .80. This yields a reference sample size (n_0) of 84.07. The actual sample size (n^*) will be the number of weights used in the distribution.

SI 6. R Session Information

```
R version 4.1.0 (2021-05-18)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 19043)

Matrix products: default

locale:
[1] LC_COLLATE=English_United States.1252
[2] LC_CTYPE=English_United States.1252
[3] LC_MONETARY=English_United States.1252
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.1252

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] kableExtra_1.3.4  EGAnet_0.9.9      patchwork_1.1.1   papaja_0.1.0.9997

loaded via a namespace (and not attached):
 [1] tidyselect_1.1.1  xfun_0.25         purrr_0.3.4       lattice_0.20-44
 [5] colorspace_2.0-2  vctrs_0.3.8       generics_0.1.0    viridisLite_0.4.0
 [9] htmltools_0.5.1.1 yaml_2.2.1        utf8_1.2.2        rlang_0.4.11
[13] pillar_1.6.2      foreign_0.8-81    glue_1.4.2        DBI_1.1.1
[17] lifecycle_1.0.0   stringr_1.4.0     munsell_0.5.0     gtable_0.3.0
[21] rvest_1.0.1       psych_2.1.6       evaluate_0.14     labeling_0.4.2
```

```
[25] knitr_1.33        psychTools_2.1.6  parallel_4.1.0    fansi_0.5.0
[29] Rcpp_1.0.7        scales_1.1.1      webshot_0.5.2     magick_2.7.3
[33] tmvnsim_1.0-2     systemfonts_1.0.2 farver_2.1.0      mnormt_2.0.2
[37] ggplot2_3.3.5     digest_0.6.27     stringi_1.7.3     bookdown_0.23
[41] dplyr_1.0.7       grid_4.1.0        tools_4.1.0       magrittr_2.0.1
[45] tibble_3.1.3      crayon_1.4.1      pkgconfig_2.0.3   ellipsis_0.3.2
[49] xml2_1.3.2        svglite_2.0.0     httr_1.4.2        assertthat_0.2.1
[53] rmarkdown_2.10    rstudioapi_0.13   R6_2.5.1          nlme_3.1-152
[57] compiler_4.1.0
```

SI 7. Interaction Plots for MBE

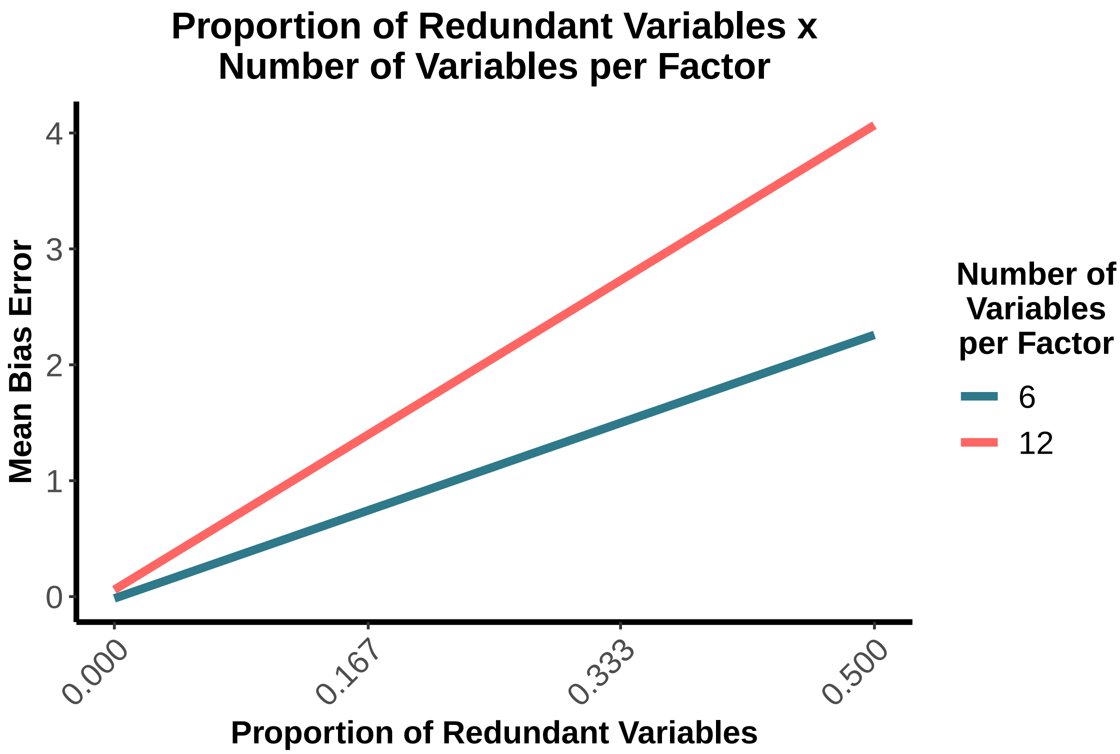

Figure 9

Mean Bias Error Interaction for Proportion of Redundant Variables and Number of Variables per Factor

SI 8. BAPQ Item Descriptions with Redundant Variables

Item	Theoretical Dimension	Item Description	Redundant
1	Aloof	I like being around other people	9
2	Pragmatic Language	I find it hard to get my words out smoothly	—
3	Rigid	I am comfortable with unexpected changes in plan	—
4	Pragmatic Language	It's hard for me to avoid getting sidetracked in conversation	32
5	Aloof	I would rather talk to people to get information than to socialize	—
6	Rigid	People have to talk me into something new	8
7	Pragmatic Language	I am "in-tune" with the other person during conversation	—
8	Rigid	I have to warm myself up to the idea of visiting an unfamiliar place	6
9	Aloof	I enjoy being in social situations	1
10	Pragmatic Language	My voice has a flat or monotone sound to it	—
11	Pragmatic Language	I feel disconnected or "out of sync" in conversations with others	—