Board 77: Exploring the Relationship between Item Stability and Item Characteristics: Exploratory Graph Analysis

Chia-Lin Tsai, University of Northern Colorado

Chia-Lin Tsai is an associate professor in the Department of Applied Statistics and Research Methods at the University of Northern Colorado. Her research interests include psychometrics studies and first-generation college students' academic experience.

Dr. Lisa Y Flores, University of Missouri, Columbia

Lisa Y. Flores, Ph.D. is a Professor of Psychological Sciences at the University of Missouri. She has expertise in the career development of Latino/as and Latino/a immigrant issues and has over 100 peer reviewed journal publications, 19 book chapters, and 3 co-e

Dr. Rachel L Navarro, University of North Dakota

Rachel L. Navarro, Ph.D. is Professor of Counseling Psychology and Associate Dean for Research and Faculty Development for the College of Education and Human Development at the University of North Dakota (UND). She is the former department chair for UNDâ€

Dr. Pat Garriott

Dr. Garriott received his PhD from the University of Missouri. He is a member of the American Psychological Association (APA), Division 17 (Counseling Psychology) of the APA, and the Society for Vocational Psychology. His work has been recognized by Divi

Han Na Suh, Georgia State University - Perimeter College
Dr. Sarah Lynn Orton P.E., University of Missouri, Columbia

Exploring the Relationship between Item Stability and Item Characteristics: Exploratory Graph Analysis

Exploring the Relationship between Item Stability and Item Characteristics: Exploratory Graph Analysis

Introduction

Motivation and Background of the Study

(This is a Methods Paper.) Scale development and validation play an essential role in research. When developing a new instrument, the literature recommends that researchers start by generating a sizeable initial item pool, guided by theory, literature, and content experts, and proceed with further evaluation and item elimination. Through a series of analyses, including item analysis, exploratory factor analysis (EFA), and confirmatory factor analysis (CFA), researchers assess dimensionality (i.e., factor structure), reduce the number of items, and then confirm the item and dimension stability in a new sample (Boateng et al., 2018). When items consistently load onto the specified dimension in a new sample (i.e., item stability), this provides evidence of an instrument's construct validity (Bowman & Goodboy, 2020).

In addition to the EFA and CFA procedures, literature suggested different item analyses, including Cronbach's alpha, item-fit (Cantó-Cerdán et al., 2021; Erhart et al., 2009), item mean, item-total correlation (Boateng et al., 2018), item information, item location, and item discrimination (Jin et al., 2018), to guide the item evaluation and reduction process. While these indices provide helpful information about item quality, less is known about how they are related to item stability (i.e., an item consistently loading onto the specified dimension in new samples). In practice, when items do not consistently load onto the specified dimension in a new sample, it is challenging for researchers to tease out whether such inconsistency is due to the item characteristics or sample characteristics (i.e., the difference between exploratory and cross-validation samples). In this study, we aim to provide some insights into *what item characteristics are related to item stability* through the newly developed exploratory graph analysis (EGA; Golino & Epskamp, 2017) and bootstrap exploratory graph analysis (bootEGA; Christensen & Golino, 2021a), which provides a way to isolate the potential confounding of sample differences. Specifically, using an engineering interest measure as an example, we explored the

relationship between item stability and the following item characteristics: 1) network loading, 2) item redundancy, 3) item mean, 4) item-total correlation, 5) item discrimination, and 6) item location. These indices were selected as they measure different aspects of item quality from different measurement frameworks, including network psychometrics, Classical Test Theory (CTT), and Item Response Theory (IRT). We explained each in more detail in the following section. In this current study, we focused on addressing the following research questions:

- RQ1: What is the relationship between item stability and the item quality indices (i.e., network loading, item redundancy, item mean, item-total correlation, item discrimination, and item location)?
- RQ2: Do stable and unstable items differ in the average value of the item quality indices?
- RQ3: How well do the item quality indices perform when classifying stable and unstable items?

Brief background on Network Psychometrics

Exploratory graph analysis (EGA; Golino & Epskamp, 2017; Golino et al., 2020) is a recently developed approach based on graph theory and psychometrics to estimate the number of dimensions underlying the multivariate data using network models. The EGA technique starts by estimating a network using the Gaussian graphical model (Epskamp et al., 2018), which captures partial correlations between items and uses the graphical least absolute shrinkage and selection operator (graphical LASSO, Friedman et al., 2008) to optimize the structure of the network. Next, the Walktrap algorithm is used to identify the number of dimensions through a sequence of partitions into communities while searching for the best organization of nodes that maximizes the modularity index (Pons & Latapy, 2006). While EGA can accurately identify the clusters of items, the generalizability and replicability of the results can be an issue due to sampling variability. Thus, the bootstrap EGA (bootEGA; Christensen & Golino, 2021a) approach was developed to evaluate the structural stability concerning dimensions and items across multiple generated samples. The bootEGA can provide information on the reproducibility and generalizability of the dimension analysis results derived from the EGA. The item stability measure is one key descriptive statistic that can be provided through EGA and bootEGA.

Item Stability. Item stability describes how often an item is placed in the same dimension, as identified by the empirical data, across multiple samples (Christensen & Golino, 2021a). Specifically, the bootEGA can generate multiple random samples based on the empirical correlation matrix from the observed data, which provides multiple samples that are consistent in the sample size and the underlying relationship between items. Thus, this item stability measure can tell researchers how often an item would load onto the specified dimension across multiple random samples with similar characteristics. This provides a great opportunity for exploring what item characteristics from the empirical data are related to item stability while isolating the potential confounding of sample differences.

Brief Background on Item Quality Indices in Network Psychometrics

Network Loading. Network loading is an important measure to represent node (item) quality in the psychometric network literature. Network loadings are formulated as each node's (item's) unique contribution to the emergence of a dimension. Network loadings provide similar information to factor loadings in the latent variable models and can be used for selecting items, testing measurement invariance, and computing factor scores (Christensen & Golino, 2021b). The standardized network loading values of 0.15, 0.25, and 0.35 represent small, medium, and large effect sizes of the loading magnitudes (Christensen & Golino, 2021b).

Item Redundancy. The concept of item redundancy is broadly defined as two items having large uniqueness correlations (Christensen et al., 2023). A measure that describes the extent to which a pair of items (nodes) overlap in the network, meaning sharing similar connections (i.e., strength, signs, and quantity) is called weighted topological overlap (wTO; Zhang & Horvath, 2005). The higher wTO values indicate greater redundancy between a pair of items. When a wTO value is greater than 0.2 for a pair of items, these items are suggested to be combined or reduced (Christensen et al., 2023).

Brief Background on the Engineering Interest Measure (EIM)

The Engineering Interest Measure (EIM) was recently developed to assess individuals' interests in engineering-related tasks and skills for the adult population in the workplace. Using exploratory graph analysis (EGA; Golino & Epskamp, 2017) and bootstrap exploratory graph analysis (bootEGA;

Christensen & Golino, 2021b), our research team examined the dimensionality of the initial EIM item pool (89 items) and further reduced the EIM to 38 items, measuring six latent constructs (Tsai et al., 2024). During the EGA and bootEGA analysis, we identified a set of 51 items that were unstable across random samples. In this current study, we explored what item characteristics are related to the stable and unstable items.

Methods

Sample

A total of 476 Latinx engineers completed all 89 items of the Engineering Interests Scale. All participants received an engineering undergraduate degree between 2015 and 2022. All were employed as engineers in the U.S. Table 1 presents the sample characteristics. The average age of the participants was 34 years old (range = 23 - 58) when the data were collected. There were 158 (33.2%) women and 313 (65.8%) men, and five participants identified as genderqueer or non-binary (1%). Most participants identified their ethnic origins as Mexican (n= 320, = 67.2%).

Table 1. Demographic Characteristics of the Sample

Variable	Count	%
Age	*Mean* = 34, *Range* = 23 – 58	
Gender		
Woman	158	33.2
Man	313	65.8
Trans man/ Trans woman/ Genderqueer/Non-binary	5	1.0
Hispanic Origin [a]		
Colombian	22	4.6
Cuban	32	6.7
Guatemalan	13	2.7
Mexican	320	67.2
Peruvian	13	2.7
Puerto Rican	14	2.9
Salvadorian	14	2.9
Venezuelan	18	3.8
Other [b]	74	15.5
Race [a]		
Black/African American	15	3.2
Indigenous American	17	3.7
Indigenous Mexican	55	11.9

Middle Eastern/North African	12	2.6
White	343	73.9
Other [c]	83	17.9
Undergraduate Engineering Major		
Civil Engineering	86	18.1
Computer Engineering	77	16.2
Electrical Engineering	76	16.0
Mechanical Engineering	121	25.4
Other	116	24.4
Work Hours Per Week		
Less than 40 Hours	18	3.8
40 Hours	239	50.2
More than 40 Hours	219	46.0

Note: N = 476. [a] Participants can select more than one category. The percentage will not add up to 100. [b] Other origins include Argentinian, Belizean, Bolivian, Brazilian, Chilean, Costa Rican, Dominican, Ecuadorian, Guyanese, Honduran, Nicaraguan, Surinamese, and not specified. [c] Other race categories include Asian, Hawaiian/Pacific Islander, Indigenous Central American, and not specified.

Item Stability Measure

In the current study, we conducted EGA on the initial item pool (89 items) to assess the dimensionality of the instrument. Next, we conducted bootEGA with 500 random samples to verify the item and structural stability of the dimensionality findings from the EGA. All the analyses were conducted using the *EGAnet* package (Hudson & Alexander, 2023) in R, and the results of the initial dimensionality assessment are summarized in Tsai et al. (2024). Based on the booEGA results, the "itemStability" function was used to generate the number of times an item is estimated in the same dimension, as initially estimated in the EGA step, across random samples. The raw item stability value is a continuous variable between 0 (0%) and 1 (100%). A binary item stability index (1 = unstable items 0 = stable items) was created based on the criteria that item stability greater than 0.75 represents an adequate number of times the item is assigned to the same dimension (Christensen & Golino, 2021a). In this study, we first used the raw item stability variable to explore an overall association pattern (RQ1) and the binary index to investigate group differences and classification accuracy (RQ2 and RQ3).

Item Quality Indices

Network Loadings. The standardized network loadings were computed from the initial EGA results for each item. In the *EGAnet* package, the "net.loads" function was used to generate the node's

strength centrality within each specified dimension, providing a similar interpretation to factor loadings (Christensen & Golino, 2021b). From the measurement perspective, items with a stronger network loading have a stronger association with the dimension and are expected to be more stable across samples.

Item Redundancy. The item redundancy measure was computed from the initial EGA results for each item (Christensen et al., 2023). In the *EGAnet* package, the "UVA" function was used to conduct the unique variable analysis (UVA), which identified locally dependent (redundant) variables in a multivariate dataset and generated the wTO measure for each pair of items (Zhang & Horvath, 2005). Because the UVA does not require prior knowledge of the dimensions, the analysis was conducted based on all the 89 items in the multivariate dataset. In this study, we computed the mean wTO for each item to represent its overall redundancy with other nodes (items) in the network.

Item Mean and Item-Total Correlation. The item mean and item-total correlation was calculated for each item within the specified dimension based on the EGA results. We used the "alpha" function in the *Psych* package to produce these item statistics. In the EIM dataset, a higher item mean indicates that participants, on average, endorsed a high level of interest in the specified engineering skill/task. A high item-total correlation represents that participants' response to a specific item is strongly associated with their total scores of the specified dimensions.

IRT Item Parameters. Based on the EGA results, the IRT item discrimination and location parameters were estimated for each item within the specified dimension. We used the *mirt* package to conduct a generalized partial credit model (GPCM; Muraki, 1992), an IRT model commonly used for modeling polytomous item responses (Dai et al., 2021; De Ayala, 2013). In our study, the item location is calculated as the average of the category locations (thresholds). In the EIM dataset, a high item location value indicates that an item requires a high level of engineering interest to endorse its categories. A high discrimination value indicates that an item can differentiate individuals' engineering interests well.

Analysis

To address our RQ1, we used Pearson correlation to analyze the association between the item quality indices and raw item stability value. The item quality indices were treated as continuous variables

(i.e., network loadings, item redundancy, item means, item-total correlation, IRT item location, and IRT item discrimination). We used the raw item stability score in the correlation analysis to explore an overall pattern of associations. A correlation value lower than 0.39 is a weak correlation, between 0.40 and 0.69 is a moderate correlation, and 0.7 or greater is a strong correlation (Schober et al., 2018).

To address our RQ2, we conducted independent-sample *t*-tests to compare mean differences in the item quality indices between stable and unstable items. The binary item stability index (1 = unstable items, 0 = stable items) was used as the independent variable. The item quality indices (i.e., network loadings, item redundancy, item means, item-total correlation, IRT item discrimination, and IRT item location) were treated as continuous variables and used as the dependent variables in the test separately. Because of the multiple testing, the risk of type-I error was greatly increased. Therefore, we used a more stringent alpha level ($p < .001$) to control for the type-I error inflation (Abdi, 2010).

To address our RQ3, we conducted separate logistic regression analyses with each item quality index as a predictor and the binary item stability index (1 = unstable items 0 = stable items) as the outcome in the model. Based on the logistic regression results, we then plotted the classification accuracy for each index using the receiver operating characteristic (ROC) curve plot. We examined the area under the curve (AUC) to determine the classification accuracy. When an AUC value equals or less than 0.5, this means a classifier is not doing better in classifying stable and unstable items than a random guess. Generally, AUC values ≥ .7 are considered acceptable, with values ≥ .8 considered excellent and values ≥ .9 considered outstanding (Mandrekar, 2010).

Findings

RQ1: Correlation between Item Stability and Item Quality Indices

The Pearson correlations of the item stability measure and the item quality indices are presented in Table 2. Our results showed that item stability had a positive, moderate correlation with network loading ($r = 0.53$), item-total correlation ($r = 0.48$), and IRT item discrimination ($r = 0.46$). Item stability had a negligible correlation with item mean and item redundancy. (rs < 0.3).

Table 2. Pearson Correlation of the Item Quality Indices

Item Quality Indices	Network Loading	Item Redundancy	Item Mean	Item-Total Correlation	IRT Item Discrim.	IRT Item Location
All 89 Items						
Network Loading	1.00					
Item Redundancy	0.22*	1.00				
Item Mean	-0.19	-0.18	1.00			
Item-Total Correlation	0.82***	0.18	-0.41***	1.00		
IRT Item Discrim.	0.84***	0.16	0.01	0.77***	1.00	
IRT Item Difficulty	0.40***	0.20	-0.92***	0.64***	0.25*	1.00
Item Stab.	0.53***	0.21	-0.18	0.48***	0.46***	0.32**

Note: 476 participants completed all 89 items (38 stable items and 51 unstable items). IRT Item Discrim: IRT item discrimination parameter. Item Stab.: raw item stability measure (continuous variable).

RQ2: Mean Differences in Item Quality Indices

The descriptive statistics of the item quality indices are presented for the stable and unstable items (Table 3). Our findings suggested that stable and unstable items differ systematically in all characteristics except for item redundancy. Specifically, stable items had significantly higher network loadings ($t = 4.36$, $df = 87$, $p < .001$), lower item means ($t = -4.03$, $df = 87$, $p < .001$), higher item-total correlations ($t = 5.03$, $df = 87$, $p < .001$), higher IRT item discrimination ($t = 3.71$, $df = 87$, $p < .001$), and higher IRT item location ($t = 5.27$, $df = 87$, $p < .001$).

Table 3. Item Quality Indices by Item Stability Status

Item Quality Indices	Stable Items (38 Items)	Unstable Items (51 Items)
	Mean (Min / Max)	Mean (Min / Max)
Network Loading	0.22 (0.12 / 0.39) [a]	0.16 (0.02 / 0.37) [a]
Item Redundancy	0.02 (0.01 / 0.07)	0.02 (0.01 / 0.07)
Item Mean	3.61 (2.88 / 4.10) [a]	3.91 (3.28/ 4.54) [a]
Item-Total Correlation	0.72 (0.54 / 0.84) [a]	0.62 (0.37 / 0.83) [a]
IRT Item Discrim.	1.50 (0.56 / 3.24) [a]	1.09 (0.38 / 2.37) [a]
IRT Item Location	-0.74 (-1.42 / 0.22) [a]	-1.32 (-2.53 /-0.43) [a]

Note: N = 476. [a] Significant differences between the two groups ($p < .001$), based on *t*-tests. IRT Item Discrim: IRT item discrimination parameter.

RQ3: Performance of Item Quality Indices

Our results suggested that several of the item quality indices can correctly classify stable and unstable items on an acceptable level. Specifically, the network loadings, item mean, item-total correlation, and IRT item location had AUC values greater than 0.7 (AUCs = 0.75, 0.72, 0.78, and 0.78, respectively). On the other hand, item redundancy (AUC = 0.68) and IRT item discrimination (AUC = 0.69) are slightly below the cut-off value.

Significance and Implications

In this study, we explored the relationship between item stability and item quality indices from different measurement frameworks through the newly developed exploratory graph analysis approach. Using an engineering interest measure as an example, our findings provide some empirical support for how different item quality indices from different measurement frameworks are related and how they are related to item stability.

First, our results showed that item stability had a positive and moderate relationship with network loadings, item-total correlation, and IRT item discrimination. This finding is consistent with previous literature that suggests that items with high item-dimension association (e.g., factor loading) play a more prominent role in the structure of the dimension. Thus, using scale items with high factor loadings to guide the item retention process should lead to a consistent conclusion in a different sample (Jin et al., 2018).

Second, in examining the descriptive statistics of the stable and unstable items, our findings showed that in comparison to unstable items, stable items, on average, were characterized by higher network loadings, higher item-total correlation, higher item discrimination as well as higher item difficulty (i.e., lower item means and higher IRT item location). Descriptive statistics revealed that for the unstable items, there are more items with a mean of four or above (on a 5-point scale) compared to the stable items. This may suggest that these engineering interest items are “easy” for many individuals to endorse high rating categories (i.e., *like* and *strongly like*), which may not provide helpful information to

differentiate participants' engineering interest levels. This finding is consistent with the literature that items with higher discrimination and greater difficulty can provide more information about a person's ability (Yang & Kao, 2014).

Our findings provide implications for engineering education researchers who want to develop a scale for measuring a latent educational or psychological construct. Particularly, during the scale validation process, when researchers observe that their items are not loading to the same specified dimension across independent EFA and CFA samples, they may consider using EGA to verify these items' behavior across multiple samples with similar characteristics to the original EFA sample. The item stability index from the bootEGA procedure could provide insights into whether the inconsistent latent structure observed between the EFA and CFA samples would be due to item or sample characteristics. When conducting EGA is not a viable option, researchers may consider using the CTT and IRT item quality indices to inform item selection, as these indices can identify unstable items to some extent.

To conclude, our findings suggested retaining items that strongly connect to the specified dimensions and items that are not too easy for individuals to endorse the high rating scale categories (e.g., *like* and *strongly like*"). Future studies may further explore the relationship between item stability and other item characteristics under different data conditions.

References

Abdi, H. (2010). Holm's sequential Bonferroni procedure. *Encyclopedia of research design*, *1*(8), 1-8.

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*. https://doi.org/10.3389/fpubh.2018.00149

Bowman, N. D., & Goodboy, A. K. (2020). Evolving considerations and empirical approaches to construct validity in communication science. *Annals of the International Communication Association*, *44*(3), 219–234. https://doi.org/10.1080/23808985.2020.1792791

Cantó-Cerdán, M., Cacho-Martínez, P., Lara-Lacárcel, F., & García-Muñoz, Á. (2021). Rasch analysis for Development and reduction of Symptom Questionnaire for visual dysfunctions (SQVD). *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-94166-9

Christensen, A. P., & Golino, H. (2021a). Estimating the stability of psychological dimensions via bootstrap exploratory graph analysis: A Monte Carlo Simulation and tutorial. *Psych*, *3*(3), 479–500. https://doi.org/10.3390/psych3030032

Christensen, A. P., & Golino, H. (2021b). On the equivalency of factor and network loadings. *Behavior research methods*, *53*(4), 1563-1580. https://doi.org/10.3758/s13428-020-01500-6

Christensen, A. P., Garrido, L. E., & Golino, H. (2023). Unique Variable Analysis: A network psychometrics method to detect local dependence. *Multivariate Behavioral Research*, 1–18. https://doi.org/10.1080/00273171.2023.2194606

Dai, S., Vo, T. T., Kehinde, O. J., He, H., Xue, Y., Demir, C., & Wang, X. (2021). Performance of polytomous IRT models with Rating Scale Data: An investigation over sample size, instrument length, and missing data. *Frontiers in Education*, *6*. https://doi.org/10.3389/feduc.2021.721963

De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.

Erhart, M., Hagquist, C., Auquier, P., Rajmil, L., Power, M., & Ravens-Sieberer, U. (2009). A comparison of Rasch item-fit and Cronbach's alpha item reduction analysis for the development of

a quality of life scale for children and adolescents. *Child: Care, Health and Development*, *36*(4), 473–484. https://doi.org/10.1111/j.1365-2214.2009.00998.x

Epskamp, S., Waldorp, L. J., Mõttus, R., & Borsboom, D. (2018). The Gaussian graphical model in cross-sectional and time-series data. *Multivariate Behavioral Research*, *53*(4), 453–480. https://doi.org/10.1080/00273171.2018.1454823

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441. https://doi.org/10.1093/biostatistics/kxm045

Golino, H., Christensen, A. P., & Moulder, R. (2020). EGAnet: Exploratory graph analysis: A framework for estimating the number of dimensions in multivariate data using network psychometrics. *R package version 0.9*, *5*.

Golino, H. F., & Epskamp, S. (2017). Exploratory graph analysis: A new approach for Estimating the number of dimensions in psychological research. *PloS One, 12*(6), e0174035. https://doi.org/10.1371/journal.pone.0174035

Hudson G., Alexander, P.C. (2023). *EGAnet: Exploratory Graph Analysis – A framework for estimating the number of dimensions in multivariate data using network psychometrics*. R package version 2.0.0.

Jin, X., Liu, G. G., Gerstein, H. C., Levine, M. A., Steeves, K., Guan, H., Li, H., & Xie, F. (2018). Item reduction and validation of the Chinese version of diabetes quality-of-life measure (DQOL). *Health and Quality of Life Outcomes*, *16*(1). https://doi.org/10.1186/s12955-018-0905-z

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, *5*(9), 1315–1316. https://doi.org/10.1097/jto.0b013e3181ec173d

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series*, *1992*(1), 1–30. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Pons, P., & Latapy, M. (2006). Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications*, *10*(2), 191–218. https://doi.org/10.7155/jgaa.00124

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, *126*(5), 1763-1768. https://doi.org/10.1213/ANE.0000000000002864

Tsai, C., Flores, L., Rachael, N., Garriott, P., Suh, H., & Hunt, H. (2024). *Investigating the Structure of the Engineering Interests Measure Using Exploratory Graph*. AERA 2024 Convention, Philadelphia, PA, United States.

Yang, F. M., & Kao, S. T. (2014). Item response theory for measurement validity. *Shanghai Archives of Psychiatry*, *26*(3). https://doi.org/10.3969/j.issn.1002-0829.2014.03.010

Zhang, B., & Horvath, S. (2005). A General Framework for Weighted Gene Co-Expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology*, *4*(1). https://doi.org/10.2202/1544-6115.1128