OPEN ACCESS

ESCOLA DE MEDICINA
PUCRS

SCIENTIA MEDICA

Scientia Medica Porto Alegre, v. 30, p. 1-8, jan.-dez. 2020
e-ISSN: 1980-6108 | ISSN-L: 1806-5562

http://dx.doi.org/10.15448/1980-6108.2020.1.37124

EDUCATION IN HEALTH SCIENCES

Assessment programs and their components: a network approach

Programas de avaliação e seus componentes: uma abordagem de rede

Jimmie Leppink[1]
orcid.org/0000-0002-8713-1374
hyjl17@hyms.ac.uk

Recebido em: 13 mar. 2020.
Aprovado em: 02 mai. 2020.
Publicado em: 19 jun. 2020.

ABSTRACT: Exams and other assessments in health science education are not random events; rather, they are part of a bigger assessment program that is constructively aligned with the intended learning outcomes at different stages of a health science curriculum. Depending on topical and temporal distance, assessments in the program are correlated with each other to a more or lesser extent. Although correlation does not equate causation, once we come to understand the correlational structure of an assessment program, we can use that information to make predictions of future performance, to consider early intervention for students who are otherwise likely to drop out, and to inform revisions in either assessment or teaching. This article demonstrates how the correlational structure of an assessment program can be represented in terms of a network, in which the assessments constitute our nodes and the degree of connectedness between any two nodes can be represented as a thicker or thinner line connecting these two nodes, depending on whether the correlation between the two assessments at hand is stronger or weaker. Implications for educational practice and further research are discussed.

KEYWORDS: Assessment; programs; connectedness; network; network analysis.

RESUMO: Exames e outras avaliações na educação em ciências da saúde não são eventos aleatórios. Ao contrário, eles fazem parte de um programa de avaliação mais amplo, alinhado construtivamente com os resultados de aprendizagem pretendidos em diferentes estágios de um currículo de ciências da saúde. Dependendo da distância local e temporal, as avaliações no programa são correlacionadas entre si em maior ou menor grau. Embora a correlação não equivalha à causalidade, uma vez que entendemos a estrutura correlacional de um programa de avaliação, podemos usar essas informações para fazer previsões de desempenho futuro, considerar intervenções precoces para estudantes com probabilidade de desistência e informar revisões em avaliação ou ensino. Este artigo demonstra como a estrutura correlacional de um programa de avaliação pode ser representada em termos de uma rede, na qual as avaliações constituem nossos nós e o grau de conexão entre dois nós pode ser representado como uma linha mais grossa ou mais fina que conecta esses dois nós, dependendo se a correlação entre as duas avaliações em questão é mais forte ou mais fraca. Implicações para a prática educacional e mais pesquisas são discutidas.

PALAVRAS-CHAVE: Avaliação; programas; conexão; rede; análise de rede.

Introduction

Curriculum developers and teachers do not have it easy. Their daily jobs are about juggling between a multitude of tasks, some of which pertain to teaching and assessment in one or several educational programs as well as the evaluation and development of these programs. Although programs do evolve over time, there ought to be a constructive alignment

[1] University of York (UY), North Yorkshire, YO, United Kingdom

between the *intended learning outcomes* at different stages of a curriculum, *what is taught* and in what ways, and *what is assessed* with which methods. Two assessments that, through topical vicinity, have a substantial overlap in the intended learning outcomes they intend to capture, will likely yield somewhat correlated results and more so if the amount of time between these assessments is relatively small (e.g., within the same academic year, or at the end of two consecutive academic years). That is, relatively better performance on one assessment tends to go together with relatively better performance on the other assessment, while relatively poor performance on one assessment tends to go together relatively poor performance on the other assessment. Absence of such a correlation may reflect a lack of reliability in at least one of the assessments, a lack of actual overlap in intended learning outcomes, at least one of the assessments suffering from limited validity due to an unintended skill influencing the results, or some combination thereof. Statistical analysis can shed light on the reliability factor and may to some degree inform a content review that will be needed to investigate the other factors.

In the light of the previously mentioned constructive alignment, assessments organized in the course of a curriculum can be conceived as nodes in a network that represents the assessment program for the curriculum at hand: the degree of connectedness of any pair of assessments can be represented as a line linking the two nodes representing these assessments, and the thickness of that line is a function of both *topical* and *temporal* vicinity [1]. That is, the more topical and/or temporal vicinity of two assessments, the stronger the correlation and therefore the thicker the line between these two assessments. While we should not mistake correlation for causation, correlations between assessments can help us to visualize and understand the correlational structure of an assessment program. This correlational structure can be used for several purposes, including (1) to make predictions of students' future performance, (2) to consider early intervention for students who are otherwise likely to drop out, and (3) to inform revisions in either assessment or teaching. Therefore, this article demonstrates how an emerging statistical method called *network analysis* [1-5] can help us in this endeavor of visualizing, understanding, and using the correlational structure of an assessment program through a simulated worked example that incorporates types of assessments and their correlations commonly encountered in educational practice. Next, this article presents a few guidelines for educational practice and future research.

Different models

A common approach to modeling correlations between assessments has been to treat different assessments as *manifest indicators* (i.e., observed variables) of so-called *latent variables* or variables that are not directly observed. In this approach, knowledge available on the part of a student is not observed directly but is assumed to be indicated by students' performance on one or more knowledge assessments that have been designed to measure that knowledge. The same holds for skills, attitudes and other traits or states of interest. For example, through their program, medical students learn several skills that are important in clinical examination, including history taking, physical examination, problem solving and patient relationship, and students' performances on clinical assessments are treated as manifest indicators of these latent skills.

If three assessments measure the same type of knowledge or skill (e.g., grammar knowledge, or probability calculus skill), core assumption in the latent variable approach is that these three assessments *commonly respond* to differences in the latent variable of interest. In practical terms, this means that students with higher degrees of that latent variable (i.e., more knowledge, or more skill) tend to score higher on these assessments than students with lower degrees of that latent variable (i.e., less knowledge, or less skill). This tendency induces a pattern of positive correlations between these assessments, with higher scores on one assessment tending to go hand in hand with higher scores on the other two assessments.

However, do we really need latent variables to explain this kind of patterns? If a group of animals – birds, cows, tigers, or other – decide to move as a group in a specific direction, that is because they communicate rather than there being some unobserved latent bird, cow, tiger or other animal to which the group of animals commonly respond. Likewise, if we have three assessments with the same types of questions, students' performances are likely going to be similar on these three assessments regardless of any kind of latent variables because of how the questions are formulated. Besides, even if assessments can be conceived as manifest indicators of latent variables like knowledge and skill, they may be indicators of several latent variables (e.g., different types of knowledge or skill being measured by the assessment) and to varying degrees from one occasion to the next.

Conceiving series of assessments in terms of networks is like a group of animals moving in the same direction; we do not need latent variables to understand correlations between assessments and how these correlations vary across years in an assessment program. Moving away from latent variables also comes with the advantage of lower demands on sample size; while a cohort of 50 to 100 students may be large enough to estimate correlations which we then visualize in a network plot, such a sample size is rarely if ever enough for meaningful latent variable modeling.

Correlations visualized in a network

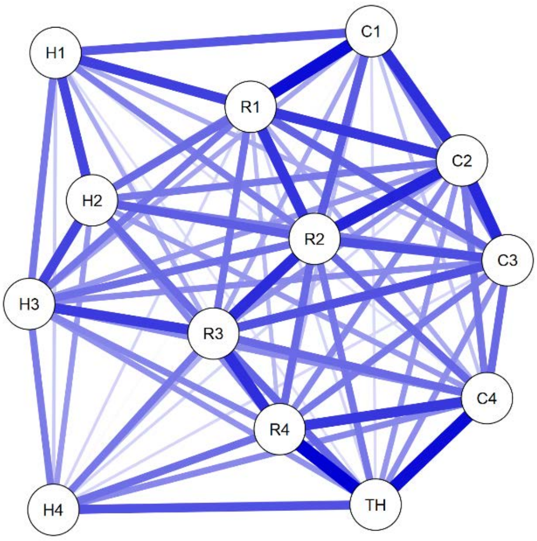
In a hypothetical Health Science Program X, which is a four-year undergraduate program, students face a total of thirteen assessments that can be categorized in four areas: *Health*, *Research*, *Communication*, and *Thesis*. For each of Health and Research, students complete a written exam towards the end of each of the four academic years. For Communication, students deliver a thirty-minutes presentation followed by fifteen minutes of questions on a predetermined topic towards the end of each of the four academic years. Finally, the thesis project runs throughout the fourth academic year and results in a written thesis at the end of the fourth year. All thirteen assessments result in a score ranging from 0% (minimum) to 100% (maximum). For the most recent cohort of *N* = 280 graduates, **Table 1** presents the correlations between these assessments (i.e., H1-H4, R1-R4, and C1-C4 represent the exams in Years 1-4 for Health, Research, and Communication, respectively, and TH stands for Thesis).

TABLE 1 – Pearson's correlations between assessments in Health Science Program X

	H1	**H2**	**H3**	**H4**	**R1**	**R2**	**R3**	**R4**	**C1**	**C2**	**C3**	**C4**
H2	0.391	---										
H3	0.287	0.395	---									
H4	0.115	0.217	0.277	---								
R1	0.403	0.315	0.288	0.094	---							
R2	0.272	0.340	0.301	0.161	0.427	---						
R3	0.180	0.313	0.413	0.285	0.295	0.444	---					
R4	0.042	0.255	0.257	0.303	0.148	0.315	0.440	---				
C1	0.357	0.293	0.191	-0.013	0.512	0.343	0.187	0.147	---			
C2	0.261	0.283	0.221	0.093	0.424	0.461	0.240	0.253	0.441	---		
C3	0.180	0.281	0.247	0.089	0.327	0.372	0.365	0.283	0.273	0.454	---	
C4	0.074	0.216	0.246	0.251	0.178	0.320	0.314	0.422	0.131	0.288	0.317	---
TH	0.072	0.168	0.233	0.368	0.091	0.286	0.322	0.538	0.082	0.221	0.238	0.513

Figure 1 visualizes the correlations presented in Table 1 in a network format (software used: JASP, version 0.11.1.0 [6], a zero-cost Open Source statistical software program that has very good facilitates for network analysis).

Figure 1 – Correlations network of the assessments in Health Science Program X: 78 non-zero edges (i.e., sparsity = 0)



In Figure 1, we only see blue lines because all correlations except for one (i.e., C1 with H4) are positive; negative correlations would be represented as red lines, and the only one out of 78 correlations that is negative is so close to zero that the line is 'lost' in the forest of blue lines. If you wonder how we got to the number of 78 correlations: given *k* variables, the number of correlations K_C that can be estimated equals:

$$K_C = [k * (k - 1)] / 2.$$

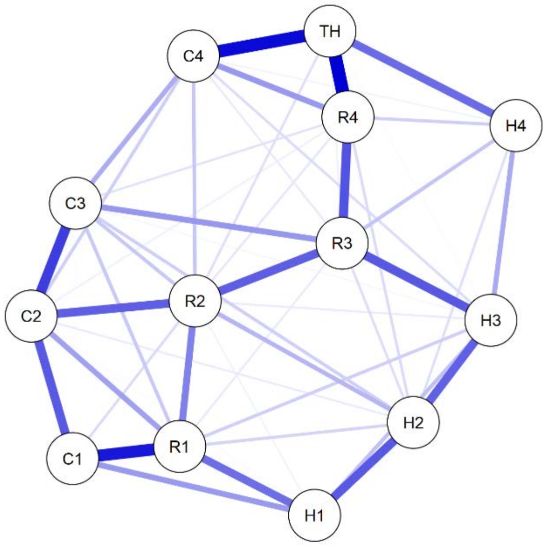
For 13 variables, this means: $K_C = [13 * 12] / 2 = 78$. We see that the correlations are strongest (Table 1) and therefore the lines are thickest (Figure 1) between adjacent exams from the same theme (e.g., H1-H2, C2-C3, R3-R4) as well as between exams from different themes taking place in the same academic year. Exams from the same theme have a high *topical* vicinity, whereas exams in the same academic year have a high *temporal* vicinity. In other words, the connectedness of any two assessments is a function of *topical* and *temporal* vicinity, which explains why correlations between exams in Year 1 and exams in Year 4 are substantially smaller (i.e., commonly in the [0; 0.2] range) than the correlations with higher topical or temporal vicinity (i.e., more commonly in the [0.3; 0.5] range).

Studying correlations between assessments that are expected to have a clear topical vicinity and that are administered in the same academic year can help to understand if that expectation of high topical vicinity is indeed realistic; correlations in the [0.3; 0.5] range or above add empirical support to that expectation. However, finding smaller correlations between assessments may point at either reliability issues for at least one of the assessments or the assessments under comparison at least partly measuring different traits or states. Reliability issues can be studied by assessing reliability statistics of each of the assessments; if these statistics indicate good reliability for each of the assessments, reliability issues are no longer a plausible explanation for the poor correlation between assessments. A similar reasoning goes for assessments that are supposed to measure the same knowledge or skill repeatedly, in the example here H1-H4, C1-C4, and R1-R4. Although a lot can happen in an academic year and therefore H2-H3 or C2-C3 correlations need not be higher than 0.5, finding correlations well below 0.3 would be surprising and again indicate either reliability issues with at least one of the assessments (to be checked through reliability statistics for each of the assessments) or the two assessments measuring quite different things.

A more parsimonious network

The correlations presented in Table 1 and visualized in Figure 1 come with a challenge: since a correlation is in practice rarely *exactly* 0 (i.e., in which case there would be no line between the two assessments at hand), it can become difficult to recognize meaningful patterns in a network – especially in the case of more assessments and hence larger networks – and it easily results in vague statements like *everything is connected with everything*, statements that are of little use to educational practice or research. Therefore, while Table 1 and Figure 1 constitute an important *starting point* to help us understand to what varying degrees different assessments are interrelated, in order to recognize important patterns more easily, we need a more parsimonious approach as a next, second step. **Figure 2** (software used: JASP, version 0.11.1.0) [6] presents such a more parsimonious approach.

Figure 2 – More parsimonious network of the assessments in Health Science Program X: 52 non-zero edges (i.e., sparsity = 1/3)



Very succinctly put, the method used to create the network in Figure 2 works as follows. The correlations presented in Table 1 are so-called *bivariate* correlations, that is: they are correlations between a given pair of assessments regardless of how strongly these two assessments correlate with other assessments in the network. Apart from bivariate correlations, we can also compute *partial* correlations: correlations between assessments' *residuals* resulting from having accounted for other assessments in the network. In a network with three assessments – A, B, and C – the partial correlation between assessments A and B is the correlation between the residuals obtained from a regression of A on C and the residuals obtained from a regression of B on C. Where more than three assessments are concerned, the partial correlation between A and B is the correlation between the residuals resulting from a regression of A on all other assessments except B and the residuals resulting from a regression of B on all other assessments except A. Although partial correlations can be weaker as well as stronger than bivariate correlations, in networks of positively correlated assessments partial correlations are usually weaker. When we then apply a technique called Least Absolute Shrinkage and Selection Operator (LASSO) [7-9], (partial) correlations that are close enough to zero shrink to exactly zero and therefore do not need to be estimated, resulting in no line connecting the two assessments under consideration. The degree to which this shrinkage takes place can be selected by using an information criterion known as the Extended Bayesian Information Criterion (EBIC) [10]. This combination of EBIC and LASSO has been called EBICglasso (e.g., the 'g' stands for 'graphical') [4-5] and is the method used to create the network in Figure 2.

In Figure 1, which uses the correlations from Table 1, all lines or 'edges' are non-zero, and therefore there are 78 correlations to be estimated. All correlations that can be estimated are estimated, and that results in a network *sparsity* of zero. In Figure 2, a total of 26 of the 78 (partial) correlations have shrunk to zero, and therefore the sparsity of the network is 26/78 or 1/3; only 52 of the 78 correlations (i.e., two-thirds) are estimated.

Different questions

The networks in Figure 1 and Figure 2 respond to different questions. On the one hand, when the question is to what degrees different pairs of assessments are correlated, we need the bivariate correlations that are visualized in Figure 1. On the other hand, when the question is which are the most important connections in a network of many assessments, Figure 2 is more useful. For example, if we are interested in predicting TH (i.e., the final assessment that is delivered in this program) performance by other assessments, Figure 1 might make one think we need all twelve other assessments in the equation, whereas from Figure 2 we learn that we probably need not look much if any further than C4, R4, and H4 (i.e., the three theme assessments in Year 4). From Table 1, we learn that a multiple linear regression model with C4, R4, and H4 as predictors explains about 41.8% of the variance in TH and that no statistically or practically significant gain in that proportion of variance explained is achieved by adding any of the other predictors. Likewise, if we are interested in the question to what extent the Year 1 theme assessments can predict R2 (i.e., the Year 2 Research assessment), Figure 2 indicates that in a regression model with C1, H1, and R1 the latter contributes most to the prediction of R2, which makes sense given the topical vicinity. From Table 1, we learn that a regression model with R1 explains about 18.2% of the variance in R2 (i.e., the square of 0.427 in Table 1), a regression model with R1 and C1 explains about 20.4% of the variance in R2, and adding H1 to the latter regression model only increases the proportion of variance explained to 21.0%, an increase which is neither practically nor statistically significant.

Examples of other questions for which network analysis can be useful

Although network analysis is in this article presented as a useful method in the context of assessment and the evaluation of an assessment program which has longitudinal themes running through the curriculum, network analysis can be used in assessment programs in which such longitudinal structures are absent as well. More broadly, network analysis is a method that has many promising applications. One line of applications is found in the context of the previously mentioned latent variable modeling. Currently, factor analytic and other latent variable methods are used to examine the psychometric structure of measurement instruments and, in some cases, how latent variables supposedly measured by different instruments may be related. Network analysis provides a complement of, or to some extent perhaps an alternative to, latent variable methods. While in latent variable models, observed variables such as questionnaire items or (in the context of this article) assessments are assumed to be causally influenced by unobserved latent variables, in network models items or assessments that measure (more or less) the same variables of interest simply cluster together in cliques in a network. Latent variable models are a bit like seeing people being positioned and moving in the same direction in response to unobserved latent people (Gods?) moving them. In network models, topical and temporal vicinity provide directly observed (manifest) variables to explain why some items or assessments (or in the metaphor: people) are more connected than others; no latent variables are needed. The thirteen assessments in the example network are exam scores, but network analysis can also be applied to for example scores of exam *sub-scores*. For example, in a clinical exam where medical students are assessed on history taking, physical examination, problem solving, and patient relationship in a series of stations each of which is one student-patient (or student-simulated patient) interaction, network analysis can help to see how the sub-scores on these different skills are intercorrelated, within and across stations (e.g., see Chapters 9-10 in [1]).

Another context in which network analysis is very useful is found in studies with repeated measurements or larger time series with the same participants; network analysis can then help to acquire an understanding of the residual variance-covariance structure of the set of measurements (e.g., Chapter 11 in Leppink [1]). Traditional, fairly simple statistical models often assume that the correlation is (more or less) constant across measurements; if this assumption is correct, the resulting network visualizing correlations (Figure 1) should be one of lines that are of more or less equal thickness (and of the same color). However, in practice, the correlation between two measurements tends to decrease with increasing temporal distance and that tends to result in patterns like the ones we see for the three longitudinal themes in Figure 1 (e.g., among R measurements, R1 is correlated most with R2 and least with R4).

Finally, in a cross-sectional context, network analysis can help to make sense of a minefield of large numbers of variables that are intercorrelated to a more or lesser extent. Without a network perspective, we might find ourselves in an exercise of very large numbers of regression models for the prediction of different variables of interest. Doing so would come at the serious risk of overlooking potentially important variables as well as of including variables in our models that do not add much to the prediction. As demonstrated in the example with Figure 2, adopting a network approach can help us to identify which are the most important predictors for any to-be-predicted variable of interest.

Guidelines for educational practice and research

As any potentially powerful statistical method, network analysis does have a cost: a substantial sample size, especially when using partial correlations (Figure 2). On a positive note, cohorts of around N = 250 can be enough for a good performance of the network approach in networks of up to 25 assessments or variables of interest otherwise [1, 11] and somewhat smaller cohorts may do for smaller networks which include clearly stronger and clearly weaker connections. For bivariate correlation networks (Figure 1), sample size is less of an issue; student cohort sizes of 50 to 100 may be fine. However, when the interest lies in the more parsimonious type of networks (Figure 2) and the cohort size becomes substantially smaller than 250 (e.g., N = 150), networks with no correlations being estimated (i.e., 100% sparsity) or an otherwise too high sparsity become more likely, and we may want to consider using the network approach for a more limited number of variables. If for example we are in a program where cohort sizes are 50 to 100 students, we may include only 5 to 10 (e.g., only R1-R4 and TH, or only the in total seven Year 3 and Year 4 assessments) assessments in our networks instead of 15 or 25 assessments. Likewise, using sub-scores instead of (as in the example) overall exam scores will require larger numbers of observations as more variables will be involved (i.e., one overall score is a combination of several sub-scores); as always, models involving more variables tend to put higher demands on sample size than models involving fewer variables. In smaller cohorts, this comes down to focusing on smaller numbers of assessments (e.g., two assessments with three or four sub-scores each instead).

Sample size limitations notwithstanding, network analysis can provide a useful supporting tool for visualizing the connectedness of assessments in a program. Interesting lines of research can be found in the study of the stability or dynamicity of networks within a program across cohorts, in the presence or absence of revisions being made to the program, and in differences between networks from different programs that have some features in common (e.g., longitudinal theme lines, or topical vicinity such as via medical or health science programs in a country or region). Further, while the example in this article – for the sake of simplicity – focuses on an assessment program, variables from student surveys about different teaching blocks or modules linked to specific assessments and about their motivational or emotional states and/or how much time they spent on different activities in a block or module can be included in the network as well to gain an understanding of the extent to which student-related and program-related factors can help to predict students' assessment performance throughout a curriculum, where we may want to consider early intervention or remediation, and where we may want to make revisions to our teaching and/or assessment.

To conclude

Network analysis is an emerging statistical approach with promising applications in a variety of contexts, including in the evaluation and revision of assessment (and teaching) programs in educational curricula. While it does *not replace* content review or other statistical methods and approaches, it can greatly *facilitate* our work as educational practitioners and researchers. Network analysis is available in zero-cost Open Source software such as JASP, and documentation for its use can be found in the list of references provided in this article.

Notes

Funding

This study did not receive financial support from external sources

Conflicts of interest disclosure

The author declares no competing interests relevant to the content of this study.

Authors' contributions.

The author declares to have made substantial contributions to the conception, or design, or acquisition, or analysis, or interpretation of data; and drafting the work or revising it critically for important intellectual content; and to approve the version to be published.

Availability of data and responsibility for the results

The authors declares to have had full access to the available data and they assume full responsibility for the integrity of these results.

REFERENCES

1. Leppink J. The art of modelling the learning process: Uniting educational research and practice. Cham: Springer; 2020. [cited 2020 Feb. 14]. Available from: https://www.springer.com/gp/book/9783030430818.

2. Borsboom D, Cramer AOJ. Network analysis: An integrative approach to the structure of psychopathology. Ann Rev Clin Psychol. 2013;9:91-121. https://doi.org/10.1146/annurev-clinpsy-050212-185608.

3. Cramer AOJ, Waldorp LJ, Van der Maas HLJ, Borsboom D. Comorbidity: A network perspective. Behav Brain Sci. 2010;33:137-50. https://doi.org/10.1017/S0140525X09991567.

4. Epskamp S, Borsboom D, Fried EI. A tutorial on regularized partial correlation networks. Psychol. Meth. 2018;23:617-34. https://doi.org/10.1037/met0000167.

5. Golino HF, Epskamp S. Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. PLoS One. 2017;12:e0174035. https://doi.org/10.1371/journal.pone.0174035.

6. Love J, Selker R, Marsman M, et al. JASP version 0.11.1.0. [cited 2020 Feb. 19]. Avalaible from: https://jasp-stats.org/.

7. Santosa P, Symes WW. Linear inversion of band-limited reflection seismograms. SIAM J Sci Stat Comp. 1986;7:1307-30. https://doi.org/10.1137/0907087.

8. Tibshirani R. Regression shrinkage and selection via the lasso. J Royal Stat Soc Ser B. 1996;58:267-88. https://doi.org/10.1111/j.1467-9868.2011.00771.x.

9. Tibshirani R. Regression shrinkage and selection via the lasso: A retrospective. J Royal Stat Soc Ser B. 2011;73:273-82. https://doi.org/10.1111/j.1467-9868.2011.00771.x.

10. Chen J, Chen Z. Extended Bayesian information criteria for model selection with large model spaces. Biometr. 2008;95:759-71. https://doi.org/10.1093/biomet/asn034.

11. Dalege J, Borsboom D, Harreveld F, Van der Maas HLJ. Network analysis on attitudes: A brief tutorial. Soc Psychol Pers Sci. 2017;8:528-37. https://doi.org/10.1177/1948550617709827.

Mailing address:

Jimmie Leppink

University of York

Y010 5DD

Heslington, York, United Kingdom