Investigation of Rater Effects Using Social Network Analysis and Exponential Random Graph Models

Educational and Psychological
Measurement

2018, Vol. 78(3) 430–459

© The Author(s) 2017

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164416689696

journals.sagepub.com/home/epm

Iasonas Lamprianou¹

Abstract

It is common practice for assessment programs to organize qualifying sessions during which the raters (often known as “markers” or “judges”) demonstrate their consistency before operational rating commences. Because of the high-stakes nature of many rating activities, the research community tends to continuously explore new methods to analyze rating data. We used simulated and empirical data from two high-stakes language assessments, to propose a new approach, based on social network analysis and exponential graph models, to evaluate the readiness of a group of raters for operational rating. The results of this innovative approach are compared with the results of a Rasch analysis, which is a well-established approach for the analysis of such data. We also demonstrate how the new approach can be practically used to investigate important research questions such as whether rater severity is stable across rating tasks. The merits of the new approach, and the consequences for practice are discussed.

Keywords

rater effects, social network analysis, exponential random graph models, Rasch model

Introduction

The quality of ratings is one of the thorniest problems faced by the psychometric community and the assessment industry. A large number of researchers across disciplines have discussed the factors governing the scores awarded by raters in high-

¹University of Cyprus, Nicosia, Cyprus

Corresponding Author:

Iasonas Lamprianou, University of Cyprus, Kallipoleos 75, Nicosia 1678, Cyprus.

Email: iasonas@ucy.ac.cy

stakes assessment contexts, such as university entrance and placement, job recruitment, or licensure examinations (e.g., Leckie & Baird, 2011; Wu & Tan, 2016). A massive body of existing literature has investigated various rating characteristics (such as severity and inconsistency of rating) and discussed how these affect negatively the measurement process and the reliability of the rating outcome (Banerji, 1999; Braun, 1988; Cushing, 1999; Engelhard, 1992, 1994, 2002; Rae & Hyland, 2001; Saal, Downey, & Lahey, 1980).

Other researchers used quantitative methods to identify personal or occupational characteristics that can help practitioners choose “better” raters. Some of the research findings are, however, contradictory, for example, Weigle (1998) found more experienced raters to be more lenient than their less experienced colleagues but Bonk and Ockey (2003) reported that returning raters (i.e., more experienced) were more severe; others, like Barrett (2001), found mixed results. It has been suggested that raters’ first language, gender, or professional training are important factors that define rater characteristics (e.g., see He, Anwyll, Glanville, & Deavall, 2013) but contradictory findings have been reported by others such as Eckes (2005), Royal-Dawson and Baird (2009), and Wiseman (2012).

The contradicting results extend the confusion and the agony of policy makers and practitioners alike who often receive much pressure by the media and the public who demand high-stakes assessment programs to have an adequate pool of competent raters in their disposal (see, e.g., BBC, 2003; Smithers, 2003). Although this discussion may sound technical, it is certainly of wider interest and has far-reaching social and political consequences in some countries. For example, the huge outcry by the public and the media after the examination fiasco of 2002 in England, forced the Chief Executive of the Council for Curriculum, Examinations and Assessment (CCEA, UK), Gavin Boyd, to state that the supply of suitably qualified raters was vital to the health of the whole examining system and that “in recent years, examiners [i.e., raters], or the lack of them, has become a major issue in education” (BBC, 2003).

Alas, identifying the profile of “the consistent rater” through quantitative methods has proved to be a chimera. For example, Congdon and McQueen (2000) used Rasch models to identify significant fluctuations in the severity of raters across a period of seven days. Hoskens and Wilson (2001) used a modified Rasch model to find that raters’ severity drifted across five successive periods. Seeking the holy grail of rater consistency, however, is not a new phenomenon. Decades ago, Freedman (1981) had suggested that severity was not a stable characteristic and could be changed through training. On the contrary, other researchers (e.g., Lunz, Stahl, & Wright, 1991) argued that rater severity was a rather stable characteristic.

Faced with public pressure and conflicting academic messages, policy makers often employ thorough (and often laborious and expensive) training procedures for the raters in high-stakes assessments (see, e.g., Masters, 2002; Myford & Wolfe, 2002). Some researchers (see Lunz, Wright, & Linacre, 1990) suggested that raters are very difficult (if at all possible) to be trained so as to have similar levels of severity. On the contrary, Stahl and Lunz (1991) suggested that training may be beneficial

in the sense that some raters may improve their consistency. Cushing (1998) also proposed that rater consistency is not a fixed characteristic but can improve through training. Barrett (2001) reported the failure of a cost-effective training package for experienced and seasonal raters, which was designed to reduce the incidence of several common rater errors. Often, qualifying sessions (also known as “moderation” or “training” sessions) are used so as to distinguish those raters who are ready to move on to operational rating from those who need further training.

In this context of contradictory and confusing results, we carried out an unpublished research (Lamprianou, 2004) which was commissioned by the English National Assessment Agency (NAA) aiming to identify examples of rating quality assurance procedures worldwide. We soon realized that Examination Boards and Testing Services alike were unwilling to talk publicly about their quality assurance procedures and we were forced to develop and distribute an anonymous questionnaire, making explicit that all responses would be treated in extreme confidentiality. Out of 26 case studies (from all continents) presented in that research, the huge majority had extensive rater training and qualifying procedures in place. The raters were asked to undergo training and in many cases they had to prove their competence by rating scripts during a qualifying session.

Because of the high demand, many quantitative approaches have been developed and validated for the analysis of rating data and are often discussed in outlets such as the *Educational and Psychological Measurement*, the *Journal of Educational Measurement*, the *Journal of Applied Measurement*, and others. Some of the most prominent techniques to analyze rating data are the Rasch model (Linacre, 1994), generalizability theory (Brennan, 2001), and multilevel models (Goldstein, 2010). All these techniques have their advantages and their disadvantages and an in-depth technical comparison goes beyond the scope of this article; however, the interested reader could refer to Baird, Hayes, Johnson, Johnson, and Lamprianou (2013) and Congdon and McQueen (2000) for more details.

In response to the recent calls, in prominent journals, for more research in this area (see, e.g., Wang, Engelhard, & Wolfe, 2016), this study uses three different datasets (i.e., two rater qualifying datasets and one simulated dataset), to demonstrate the use of an intuitive, visual, and quantitative approach, to explore the readiness of a group of raters to proceed to operational rating. More specifically, social network analysis (SNA) and exponential random graph models (ERGM) are used to analyze a simple simulated dataset and the results are presented in a didactic style. Then, empirical data from two rater qualifying sessions are analyzed using our proposed SNA/ERGM approach and the results are compared to those of a Rasch analysis, which is a more “traditional,” well-established and well-known method for the investigation of rater effects.

A Conceptual Approach to SNA for Rater Effects

SNA has been used for many years in order to model relationships between individuals, events, or other entities (see Snijders, Van de Bunt, & Steglich, 2010). A node, (often known as vertex), represents an actor and an edge, (often known as link),

Figure 1. A triad of raters who rated the same examinee scripts.

between two nodes represents the relationship that connects them. Figure 1 represents the relationship of a triad of actors, A, B, and C. The link between the nodes could represent any kind of relationship; in our context, the triad of nodes could represent three raters who have rated the same examinee scripts.

Depending on the nature of the study, one may generate SNA graphs (like Figure 1) using various rules. If, for example, the aim of the study is to evaluate rater disagreement, then, an edge (i.e., a “link,” or else, a “tie”) between Rater A and Rater B could suggest that there was a substantial disagreement (i.e., a “discrepancy”) between their ratings. In practice, directed networks may be used to show who, among the raters, was the most severe or the most lenient. For example, an arrow emerging from Rater A and ending at Rater B could suggest that Rater A had awarded a significantly more generous rating compared with Rater B. However, to make our discussion more realistic, we would need to accommodate for the fuzziness and the unreliability of human behavior. For example, double-headed arrows (i.e., “mutual” links or “reciprocity” in the language of SNA) suggest that Rater A is sometimes significantly more generous and some other times significantly stricter compared with Rater B. The extent of reciprocity in the data is an indication of low reliability, although a small amount of reciprocity would always be expected in real life. In the literature of rater effects, in the case of intense reciprocity, we would say that the raters were inconsistent (i.e., there is low interrater reliability), in the sense that they were appearing to be more or less lenient almost haphazardly. In empirical settings, we normally wish leniency to be stable, at least across short time intervals (Lamprianou, 2006).

Figure 2. A directed network of five raters.

Figure 2 is an example of a directed network illustrating the disagreement between the ratings of a set of five raters. An arrow from Rater D to Rater B ($D \rightarrow B$) means that Rater D is more lenient. Rater C, however, is sometimes more lenient and sometimes less lenient, compared with Rater D; thus, the relationship between C and D demonstrates some reciprocity. In the psychometric literature, reciprocity has been quantified using various techniques, for example, in the context of Rasch models, fit statistics such as the infit and the outfit mean squares have been heavily used (see, e.g., Lamprianou 2006, 2008).

In real life, we often wish to quantify the extent to which a rater tends to disagree with the rest of the group of raters. To delve a little bit deeper in the SNA parlance, Rater B, in Figure 2, has a “degree” of 2 because there are two arrows linking that rater with other raters. Rater B has been strict but consistent, in the sense that he or she is a net receiver of edges (there are no mutual links). Rater C is more lenient than Rater B but very inconsistent; he or she has mutual links with both Raters A and D. The fact that Rater C is unpredictable is very inconvenient, however, because, if we had wished to adjust his/her ratings for severity, that would be difficult (Wu and Tan, 2016).

In the network graph parlance, rater E is an “isolate” (Figure 2, upper right corner), in the sense that he or she is not linked to any of the other raters. In disagreement networks, ideally all raters would be isolates, as we would normally wish our raters to have no major disagreements between them. Note, however, that if we were modelling agreement rather than disagreement relationships, we would wish to have no isolates at all in the network.

In addition to the relationships between pairs of raters, we are often interested in the relationships of larger groups of raters, such as triads. In fact, there are two fundamental arrangements of triads in SNA theory that are especially relevant in the study of rater effects: cyclic triads and transitive triads. A cyclic triad is a specific three-cycle arrangement where $A \rightarrow B$, $B \rightarrow C$, $C \rightarrow A$ (Block, 2015, p. 164, Figure 1b). This implies that while A is more lenient than B and B is more lenient than C, C appears to be more lenient than A! This is a counterintuitive arrangement in rating data which can only happen if there are significant inconsistencies between raters (we will investigate an example later); thus, any rating dataset needs to be investigated for the existence of cyclic triads.

Contrary to cyclic triads, the arrangement of weak transitivity demands that $A \rightarrow B$, $B \rightarrow C$, $A \rightarrow C$ (Block, 2015, p. 164, Figure 1a). In other words, if Rater A is more lenient than Rater B, and B is more lenient than C, then A will also be more lenient than C. Weak transitivity is a very desirable phenomenon in rating data, and we would expect it to be prevalent in every high-stakes rating dataset. It would also be possible to investigate for strong transitivity effects, which requires that $A \rightarrow C \rightarrow A \rightarrow B$, $B \rightarrow C$, which is a much stronger requirement for rater data. Again, these requirements need to be framed so that they account for the random error which usually exists in empirical data.

In addition to the above, assortativity is an important measure which characterizes the underlying mechanism of the formation of links between the nodes of a network. Assortativity is the extent to which nodes of high degree tend to form links with other nodes of high degree (Williams & Del Genio, 2014). On the other hand, disassortativity is the extent to which nodes of high degree form links with nodes of low degree. This can be a very important concept in the analysis of rating networks in the sense that disassortativity suggests that the raters who tend to disagree frequently with the others, tend to do so even with raters who overall demonstrate very low rates of disagreement. This could be an indication that there are raters who tend to disagree with the others in an almost haphazard way.

Unfortunately, assortativity is a rather underresearched concept, which gets increasingly more complex while we move from undirected networks, into the world of directed networks. Acknowledging this lack of research, Foster, Foster, Grassberger, and Paczuski (2010) discuss the need to investigate specific directed assortativity measures such as $r(out, in)$ and $r(in, in)$. These are assortativity measures where the first element in the parentheses indicates the degree of the node receiving an edge and the second indicates the degree of the node sending an edge. In other words, in a disagreement network, $r(out, in)$ summarizes the pattern of nodes

with high in-degree (i.e., more severe raters) to disagree with nodes with high out-degree (i.e., more lenient raters). Although such a phenomenon is to be expected, a high $r(in, in)$ assortativity measure would need to be scrutinized in the sense that we would not expect severe raters to disagree frequently with other severe raters. However, as it was discussed in previous paragraphs, choosing which assortativity measure to use in a practical setting can be heavily context dependent.

When considering larger groups of raters, under certain conditions, segregation could mean that two or more groups of raters could be drifting away (Hoskens & Wilson, 2001), so that they are rating in particularly different ways. If we model agreement (instead of disagreement) networks, then segregation would mean that the raters tend to agree within the group disproportionately frequently (and appear not to agree with raters from other groups). This could typically happen when a group of raters train together and rate in isolation from other groups of raters, so they gradually develop their own microculture and rating techniques. In such a case, we might say that the raters demonstrate a kind of “homophily” (McPherson, Smith-Lovin, & Cook, 2001) because they disproportionately tend to form links with raters with the same characteristics.

A Mathematical Approach to SNA for Rater Effects

Adapting the notation of Robins, Pattison, Kalish, and Lusher (2007), let i and j be raters of a set N of n raters. Let us assume a random variable Y_{ij} where $Y_{ij} = 1$ if there is a directed network edge from rater i to rater j (otherwise, $Y_{ij} = 0$). In our context, an edge might represent either agreement or disagreement, depending on whether we want to construct agreement or disagreement networks. Let y_{ij} be the realization (an observation) of the variable Y_{ij} and let $\mathbf{Y}$ be the matrix of all variables with $\mathbf{y}$ the matrix of observed edges. Note that in the case of an undirected network, $Y_{ij} = Y_{ji}$. In an undirected disagreement network, $Y_{ij} = 1$ means “raters i and j disagree” whereas in a directed network $Y_{ij} = 1$ means “raters i and j disagree and rater i is more lenient than rater j .”

Fundamental network statistics such as degree may be computed with simple mathematical operations over y_{ij} . For example, in an undirected network, the degree of rater i is simply the sum of all y_{ij} for all $n - 1$ raters (i.e., it is the number of other raters with whom rater i disagrees). However, for directed networks, one might consider separately the in-degree and out-degree for each rater. Out-degree is the degree of outgoing edges (a proxy of leniency) whereas in-degree is the degree of incoming edges (a proxy of severity). For example, in-degree and out-degree may be formulated as follows:

$$\text{Out-degree for rater } i : \sum_j y_{ij} \quad (1)$$

$$\text{In-degree for rater } i : \sum_j y_{ji} \quad (2)$$

At the level of the whole network, one might be interested in statistics such as the density, which is simply the proportion of all possible ties that are realized in the network and could easily be computed (in the general case of undirected networks) as

$$D(Y) = \frac{1}{\binom{n}{2}} \sum_{i < j} y_{ij} \quad (3)$$

Based on Block (2015) but adapting for the notation introduced above, we can express some fundamental concepts of a directed network as functions of in-degree and out-degree for each rater. For example,

$$\text{Reciprocity involving rater } i : \sum_j y_{ij} y_{ji} \quad (4)$$

$$\text{Transitive triplets involving rater } i : \sum_{j, h} y_{ij} y_{ih} y_{hj} \quad (5)$$

$$\text{Three-cycles involving rater } i : \sum_{j, h} y_{ij} y_{hi} y_{jh} \quad (6)$$

However, although handy, presenting formulae for all the useful statistics goes beyond the scope of this article. Most of the statistics we refer to in this article are easy to conceptualize and compute and the interested reader may refer to relevant articles such as Block (2015) or other articles we cited above. A long introductory or technical presentation to the specific method is beyond our scope; here, we aim to show how the method may be conceptualized and applied for the analysis of rater effects. Our approach is that there is no need to develop new terminology or new methods because these are already available.

An ERGM Approach for Rater Effects

All the discussion we had above, regarding the analysis of the rater data using SNAs, considers a specific realization of all the possible rating datasets we could have obtained. The data in hand, however, are just a specific observation, a snapshot in time. The same group of raters would have certainly behaved differently, and a different dataset would have been obtained if we were to repeat the rating task again. Thus, we need a stochastic model which will take into account randomness as well as structurally missing data; ERGMs allow us to draw statistical inference regarding the processes of network formation (i.e., how our data were generated).

ERGMs have been developed to analyze graph data and have been applied in various disciplines and settings (Goodreau, 2007; Robins et al., 2007). They are exceptionally appropriate to study concepts and measures of interest, such as reciprocity and cyclic triads, as they have been discussed above. It is only reasonable to assume that, in rating datasets, links (e.g., disagreements) do not happen randomly, but follow specific configurations: for example, if Rater A is more lenient than Rater B, and Rater B is more lenient than Rater C, then we would normally expect Rater A to be

more lenient than Rater C (i.e., this is the phenomenon of weak transitivity that we discussed above). Quantities to represent such effects governing the formulation of links between the nodes are called parameters. Extending the mathematical notation introduced above and drawing from Robins et al. (2007, pp. 178-179), an ERGM might be:

$$\Pr(\mathbf{Y}=\mathbf{y})=\left(\frac{1}{\kappa}\right)\exp\left\{\sum_A\eta_A g_A(\mathbf{y})\right\} \quad (7)$$

where

- A represents all configurations of the network,
- η_A is a parameter of configuration A
- $g_A(\mathbf{y}) = \prod_{y_{ij} \in A} y_{ij}$ is a network statistic
- $g_A(\mathbf{y}) = 1$ means that configuration A is observed in the network $\mathbf{y}$; otherwise $g_A(\mathbf{y}) = 0$
- κ is a normalizing quantity

The above formulation allows us to investigate the determinants of link formation, that is, in other words, we can investigate whether parameters of interest (e.g., reciprocity), appear in our data in inflated prevalence, beyond mere randomness. Using ERGMs, we can answer questions such as “if we generate a large number of random networks, with the same characteristics (e.g., the same number of raters), would we observe the same degree of reciprocity as has been observed in our data?” If not, we may infer that there is too much reciprocity in our data, which means that the raters demonstrate a larger, than randomly expected, degree of inconsistency. In a practical setting, policy-makers would need to reflect seriously on such a situation and decisions might probably need to be taken (i.e., raters to be retrained or refrain from operational rating).

Established Approaches to Analyze Rater Data

Some of the most prominent techniques used in the past to analyze rater data are the Rasch model (Linacre, 1994), generalizability theory (Brennan, 2001) and multilevel models (Goldstein, 2010). Baird et al. (2013) compared the three methods and found that, for their dataset, they produced similar results; however, the authors highlighted some discrepancies and discussed the different theoretical and historical roots between the methods (also see Engelhard, 2013, for more details).

For purposes of brevity, we present the Rasch model only in the online Appendix—Part A. It is important, though, to mention that some of the older Rasch literature is related to earlier attempts to use graph theory to represent “. . . assessment network[s] . . . as a connected system of rater and task banks for large-scale performance assessments” (Engelhard, 1997, p. 20). Although that literature

did not actually go beyond the mere description of different data collection designs for the investigation of rater effects through many-facets Rasch models, it provided important conceptualizations of complete, incomplete and nonlinked assessment networks.¹

Throughout the article, one may observe that there are many conceptual similarities between the SNA/ERGM approach and past uses of Rasch models to study rater effects. For example, as discussed above, the Rasch fit statistics may be conceptualized as analogous to reciprocity and the Rasch severity parameter may be conceptualized as analogous to the information conveyed by the in-degree and out-degree statistics. To facilitate interpretations of how research questions about rater effects can be addressed using the main SNA/ERGM concepts, see Table 1.

A Didactic Presentation of the SNA/ERGM Approach on a Simulated Dataset

Let us simulate a very simple scenario where 20 raters are invited to judge whether a number of candidate CVs (curriculum vitae) are appropriate to be considered for a specific job. Each of the raters is presented with 100 CVs. For each CV, each rater awards a rating of “1” (= the candidate is appropriate for the job) or “0” (= the candidate is *not* appropriate for the job).

Let us assume that each of the CVs to be judged has an inherent property θ , which indicates it’s “true” (as we would call it in the psychometric literature), appropriateness for the job. Let $\theta \sim U(-5, 5)$; any other distribution would do. Let us also assume that each of the raters has an inherent severity property δ drawn from a distribution $\delta \sim U(-0.5, 0.5)$; as before, any other distribution would do. More severe raters are less likely to consider a CV to be appropriate for the job. For the sake of simplicity and according to common practice, let θ and δ be stable quantities during the rating process (e.g., a CV does not change during the process and a rater’s severity also does not change). Any set of values (i.e., distributions) could have been chosen either for the CV or for the rater parameters and this would not change the generality of this example.

Let us assume that the rating outcome is governed by a standard and well-known psychometric model, such as the simple Rasch model. Let the probability of a rater with severity δ to judge a CV with appropriateness θ as appropriate for the job (i.e., outcome = 1 rather than 0) to be described by

$$P(X = 1 | \delta, \theta) = \frac{e^{\theta - \delta}}{1 + e^{\theta - \delta}} \quad (8)$$

As a consequence, the outcome of the rating process is a rectangular matrix with 100 rows (i.e., one for every CV) and twenty columns (i.e. one for every rater) where the entries are 0s and 1s. Note that the matrix could be transposed without altering the generality of this example. Since this is a fully crossed design (i.e., all raters rated all CVs), it is possible to convert this matrix to a square matrix with 20 rows and 20

Table 1. A Summary of Social Network Analysis (SNA) Concepts and Relevant Research Questions.

SNA concepts	Practical question addressed (for a disagreement network)	Comments
Isolates	Are there any raters who appear to never disagree with other raters?	This concept is useful to identify raters who tend not to disagree with other raters. This is an indication that the rater is not of extreme severity (either too lenient or too severe) because this would cause disagreements with other raters. It also means that the rater is rather consistent (i.e., a stable severity). In a Rasch analysis, this measure conveys information related both to the rater estimate but also to rater fit statistics.
In-degree (or out-degree)	Are there any raters who appear to be too severe (or too lenient)?	These two concepts (in-degree and out-degree), although useful, do not give any indications about the inter- or intra-rater consistency of the rater. The information they convey is similar to the rater severity estimates produced by a Rasch analysis.
Reciprocity	Are there any raters who sometimes appear to be more lenient and sometimes more severe?	Reciprocity may be conceptualized as an index of inconsistency. Raters demonstrating a high degree of reciprocity produce both out-going and in-coming edges with the same raters, thus appearing sometimes more lenient and sometimes more severe. Reciprocity conveys information similar to that conveyed by rater fit statistics in a Rasch analysis. Because reciprocity considers only edges produced between a pair of raters, it is not possible to know if Rater A, Rater B, or both are the source of the disagreements.
Cyclic triads	Are there any counter-intuitive patterns in the data, such as while A is more lenient than B, and B is more lenient than C, sometimes C appears to be more lenient than A?	This implies that there may be reciprocity between Raters A and C. Cyclic triads, however, consider subnetworks of three raters rather than pairs of raters. Diagnostically, this measure can help practitioners to identify raters who are possible sources of disagreements.

(continued)

Table 1. (continued)

SNA concepts	Practical question addressed (for a disagreement network)	Comments
Transitivity (weak and strong)	Are there triads of raters who demonstrate consistent rating patterns of increasing severity?	Transitivity suggests that, although there are three raters who disagree between them, there is a stable pattern of increasing severity between them. Similarly to cyclic triads, transitivity, considers subnetworks of three raters; contrary to cyclic triads, transitivity can be a positive health signal in the rating dataset.
Assortativity	Is there a correlation between the degree measures (in-degree and out-degree) of raters?	Assortativity can help identify undesirable patterns in the data such as high $r(in, in)$ correlations. This could suggest the presence of reciprocity patterns in the data.
Homophily and segregation	Are raters' background characteristics (e.g., gender or experience) affecting the probability to observe disagreements between them?	These concepts can be diagnostically very useful, especially at early stages of rater training. In a typical Rasch analysis, this question could be partly investigated by comparing the rater severity or the fit statistics of different groups of raters.

Figure 3. The distribution of disagreements between all possible pairs of raters in a simulated dataset.

columns, where both rows and columns represent raters and the intersections of rows and columns represent all the possible pairs of raters. The entries of the square matrix are y_{ij} (i.e., for every combination of raters i and j), and let them represent the count of disagreements where rater i awarded a higher rating compared to rater j . Thus, y_{ij} can take integer values $0 \leq y_{ij} \leq 100$ where higher values mean that rater i is more lenient than rater j . For example, $y_{ij} = 10$ means that for 10 CVs, rater i gave a rating of 1 whereas rater j gave a rating of 0. This is a weighted, (or “valued”), directed disagreement matrix, where the diagonal may be left blank (there is no practical gain in comparing the raters with themselves). Also note that y_{ij} is not necessarily the same as y_{ji} in directed networks.

Out of 380 ($=20 \times 20 - 20$ entries of the diagonal) possible pairs of raters i, j (i.e., $i \neq j$), the distribution of y_{ji} for the simulated data is illustrated in Figure 3. Most of the pairs of raters seem to disagree about ten times. However, there are some pairs of raters who experienced larger disagreements, for example, there are pairs of raters who disagreed 15 times or more out of 100 CVs. Note that the agreement/disagreement rate is sensitive to the choice of distributions for θ s and δ s; however, the particular choices do not affect the didactic nature or the generality of this example.

Since $0 \leq y_{ij} \leq 100$, we would need a valued network to analyze the data. However, to simplify this introductory example, we dichotomize y_{ij} so as to produce a binary (instead of valued), directed, disagreement matrix by creating a link $i \rightarrow j$ ($y_{ij}^* = 1$) in the cases where originally $y_{ij} > 10$ (otherwise, $y_{ij}^* = 0$).

Figure 4 presents the binary directed disagreement network between the 20 raters. The larger the absolute value of the severity of each rater, the larger the size of the vertex. The circles indicate raters with IDs ranging from 1 to 10, who were simulated to have negative severity (i.e., the first 10 simulated raters are more lenient than the last 10). The rectangles (IDs ranging from 11 to 20), indicate more severe raters. The

Figure 4. A social network analysis (SNA) representation of the simulated dataset. Circles represent raters with a negative sign of severity parameter (i.e., lenient raters) and squares represent raters with positive severity (i.e., severe raters). The size of the node represents the magnitude of the severity (e.g., a larger circle represents a more lenient rater).

mean out-degree is 10.8 ($SD = 6.2$) and the mean in-degree is also 10.8 (by definition), although the standard deviation for the in-degree is smaller ($SD = 4.8$). As expected, one can easily observe that the more severe raters (those with the larger IDs) have a higher number of in-degree statistics whereas the more lenient raters (those with the smaller IDs) have larger out-degree statistics.

An interesting observation is that there are no cyclic triads in the network (see previous discussion) but this is not a surprise because we have simulated the data to conform to the Rasch model. On the other hand, there are three raters who demonstrate reciprocity: each of raters with IDs 5, 9, and 11 have five mutual links with other raters. Some degree of reciprocity (=randomness) in the data is expected due to the probabilistic nature of the Rasch model. To investigate whether reciprocity is too prevalent in our data (i.e., to establish some inference regarding the data generation mechanism), an ERGM can be used. For the sake of brevity, the actual R code, the output and some commentary are presented only in the online Appendix—Part B (we present the full software output as done by Harris, 2014).

The negative sign of the coefficient of mutuality, and the small standard error, suggest that there is too little mutuality in the data (less than what randomness would command). It is thus possible to generalize our finding beyond the observed dataset

Figure 5. The absolute difference of the probability for a positive judgment by two raters ($p_1 - p_2$), as a function of the difference between their severity measures.

and suggest that the data generation mechanism does not favor reciprocity. When analyzing empirical data, such a finding would be an important health signal in the sense that we aim to investigate whether a group of raters is ready to proceed to operational rating.

It is only reasonable to assume that there is a relationship between the simulated Rasch parameters and the probability of disagreements between the raters. Reworking Equation (8), the log-odds for rater r_1 with severity δ_1 and probability p_1 to judge a certain CV with appropriateness θ as being appropriate for the job ($p_1 = 1$ rather than 0) is given by

$$\log\text{-odds } \frac{p_1}{1 - p_1} = \theta - \delta_1$$

Similarly, for rater r_2 with severity δ_2 ,

$$\log\text{-odds } \frac{p_2}{1 - p_2} = \theta - \delta_2$$

The difference in the log-odds of the two raters to give a positive judgment for a given CV is not a function of the θ parameter, so it is not strictly related to the CV; instead it is governed only by the δ parameters. The difference in the log-odds increases while the quantity $\delta_2 - \delta_1$ increases, thus,

$$\Delta(\log\text{-odds}_1 - \log\text{-odds}_2) = \delta_2 - \delta_1$$

Visually, the absolute difference between the probabilities p_1 and p_2 to report a positive judgment for any CV, as a function of the difference of the severity measures, is illustrated by Figure 5. Therefore, we may easily deduce that raters whose severity

measure is near the average severity measure of the group of raters will have a smaller probability to disagree significantly with the rest of the raters and thus to form in-degrees and out-degrees.

Summing over all possible pairs of raters, we would expect that the out-degree and the in-degree of rater r_1 would be a function of the difference between δ_1 and the average severity of the rest of the raters (which is practically the mean severity of the group, at least for large groups). Thus, we would expect a significant correlation between the simulated Rasch severity parameters and the in-degree and out-degree statistics of the raters.

In fact, for the simulated data, the correlation between out-degree and the simulated Rasch severity is $r = -0.58$ ($p < .01$) and the correlation between in-degree and simulated Rasch severity is $r = 0.61$ ($p < .01$).

Finally, we analyzed the simulated data with the Rasch model in order to recover the rater severity parameters (this is a standard practice in the psychometric literature when one wishes to investigate how well a model can recover the simulated values—see Wang & Chen, 2005).

The correlation between the Rasch recovered parameters and the out-degree (or in-degree) is close to 1 ($r = 0.94$), which suggests that analyzing the simulated data with the SNA/ERGM approach rather than the Rasch model does not lead to significant loss of information even though the data were simulated using the Rasch model in the first place! In addition to this, the SNA mutuality is also highly correlated to the Rasch infit mean square statistic ($r = 0.72$, $p < .001$), which is a standard measure of inconsistency in the established psychometric literature (see Lamprianou & Boyle, 2004, for a practical example).

In this section, we followed a didactic approach to analyze a simulated dataset using the proposed SNA/ERGM approach. To help practitioners and policy makers, we compared our findings to those of the Rasch model, which is a psychometric model for the analysis of rater effects, well known to the assessment industry and research community alike. Capitalizing on the analysis of the simulated data, the next section presents a typology of the desirable and undesirable characteristics of a rating dataset, using SNA terminology.

A Typology of Desirable and Undesirable Characteristics of a Rating Dataset

To make our article more practically useful both for researchers and practitioners, we suggest a typology of desirable and undesirable characteristics that every rating dataset should have (or not have). We focus on the fundamental SNA concepts developed in previous sections, but we also introduce a few other, secondary, concepts.

Adapting the SNA terminology for a rater disagreement network, we suggest the following:

1. The network should present a very large proportion of isolates, that is, raters who tend to agree with all the other raters.
2. Raters who are not isolates, should have low degree measures (both in-degree and out-degree), that is, they should have disagreements only with a very small number of raters.
3. Raters who have very high in-degree or out-degree need to be identified as early as possible and their ratings need to be checked (at least a sample of them). These raters may need to be removed from the rating task or be retrained.
4. The network should present negligible levels of reciprocity, which means that there should not be an increased likelihood for some raters to “send” links to those from whom they “receive” links. No matter what the severity of a rater is, it should remain stable compared with the severity of the other raters. In other words, if Rater A tends to award higher ratings compared with Rater B, he or she should consistently do so for the duration of the rating exercise (within some predetermined tolerance level).
5. The network should present negligible levels of a specific three-cycle arrangement, known as cyclic triad. If Rater A awards higher ratings compared with Rater B, and if Rater B awards higher ratings compared with Rater C, then Rater C should *not* award higher scores compared with Rater A.
6. The network should present very high levels of weak transitivity (i.e., the proportion of triads where $A \rightarrow B, B \rightarrow C \Rightarrow A \rightarrow C$) and strong transitivity (i.e., the proportion of triads where $A \rightarrow C \Rightarrow A \rightarrow B, B \rightarrow C$). High levels of transitivity is a health indicator which may be conceptualized as an antagonistic phenomenon to the cyclic triad patterns explained above.
7. There should be a large and negative correlation between the in-degree and out-degree statistics of the raters. In other words, nodes (i.e., raters) who tend to produce out-going edges (who are more lenient than others) should be considerably less likely to produce incoming edges (to be more severe than others).
8. Some assortativity measures of the network need to be routinely investigated in order to determine whether more severe raters tend to disagree with the more lenient raters, or whether disagreements are more widespread in the network, even between raters with modest leniency. Depending on the context and on the values of assortativity measures for a specific network, different strategies to retrain the raters (or revise the rating rubrics) may be needed.
9. If agreement networks are used, homophily and segregation can be important.
 - a. the network should demonstrate negligible levels of homophily.
 - b. if some degree of homophily exists, it should be investigated whether it relates to specific rater characteristics because this would be an indication of a systematic inconsistency between groups of raters (e.g., more experienced vs less experienced);

- c. the network should present negligible levels of segregation (see Bojanowski and Corten, 2014); that is, groups of raters who trained in isolation should not exhibit different rating characteristics compared to other groups of raters (it can happen in extensively de-centralized systems);
10. While time passes, the consistency of a group of raters may gradually change, as raters retire and new raters join the group. The fundamental network characteristics (e.g., transitivity, assortativity, reciprocity, homophily, proportion of isolates, mean degree, etc.) need to be monitored longitudinally, in order to identify significant shifts relating to the rating characteristics of the group.

The above suggestions are general enough for practitioners to adapt in their own context, but are not exhaustive. There is always space for further empirical guidelines to be proposed for specialized rating contexts, however, it makes more sense to restrict ourselves to some fundamental rules for the sake of generality. While researchers and practitioners familiarize with the SNA/ERGM approach to the analysis of rating data, it is likely that more detailed or specialized rules of thumb may be put forward in the future. The next sections demonstrate how some of these suggestions may be implemented in practice, for the analysis of two empirical datasets. Since these guidelines are heavily context-dependent, we first discuss the context in which the empirical data were collected.

The Background of the Empirical Data

This research uses data from a university entrance, high-stakes and very competitive examination. In this dataset, the raters rated scripts from two different language tests, so we have two empirical datasets, see Table 2.

Test A consisted of questions measuring linguistic ability (e.g., “explain the meaning of these words,” “use these words properly to construct meaningful sentences,” “give the summary of this passage,” etc.). The expected responses were sentences or short passages of free text. According to the rubrics, the raters were instructed to focus their attention on spelling, grammar, structure, as well as the content (i.e., the ideas) of the responses. The rubrics were analytic, that is, they explained in detail how many marks should be awarded on each question for different types of responses.

Test B was a long essay question and the examinees were asked to elaborate and take a position on a controversial topic that was related to Citizenship Education and its effect on the society (the essay was expected to be a few pages long). The raters evaluated the essays using six criteria such as content, language, and structure. The language criterion focused on spelling, the richness of the vocabulary and the clarity and style of writing. The content criterion focused on the appropriateness and originality of the ideas of the essay and the adequacy of the arguments. The structure

Table 2. Background Information for Tests A and B.

	Test A	Test B
Test content	Reading comprehension, grammar, and spelling	One long essay
Structure	12 questions with a total of 70 marks (6, 5, 5, 4, 2, 2, 2, 2, 2, 20, 12, 8)	6 different criteria with a total of 30 marks (7, 5, 4, 6, 3, 5)
Qualifying rating	10 scripts (from 10 candidates)	10 scripts (from 10 candidates)
No. of raters	59	59

criterion evaluated the appropriateness of the structure of the paragraphs and the structure of the essay in general (i.e., the essay should be separated in the introduction, the body of the essay and the conclusion).

Examples of high-quality responses were provided and were discussed with the raters during a training session; the raters were free to use their judgement to award marks in-between two levels, for example, if a specific response was perceived by the rater to be halfway between level 5 (10 marks) and level 4 (8 marks), he/she was allowed to award 9 marks.

The raters were secondary education, language teachers and many of them were returning raters (i.e., they had recently rated similar scripts for this examination at least once before). No information about their demographic characteristics besides years of teaching experience were made available for purposes of confidentiality.

Before the rating, the raters were trained. Detailed rubrics were given to them which had been prepared by the team of chief examiners who had authored the tests. All the raters were physically present in the same room and they discussed the rubrics one by one. When the group felt that they had reached an adequate level of consistency, each of the raters was given ten scripts to rate. In this study, we use the rating data of those ten scripts. It is important to mention that we use data from the rater qualifying phase, so it is normal to explore the data to decide if the group of raters has reached a high agreement rate so as to start operational rating.

SNA/ERGM Analysis of Empirical Data

Recording Discrepancies Between the Raters

In order to qualify for the operational rating, all 59 raters rated the same 10 scripts for Test A and the same 10 scripts for Test B (so a total of 20 scripts were used in the qualifying session). For each script, the raters awarded a score between 0 and 70 for Test A and between 0 and 30 for Test B. The ratings awarded by each of the raters were compared with the ratings of all the other raters, one pair of raters at a time. According to the official regulations of the organization that provided the data, whenever the ratings of any pair of raters disagreed by more than 10% on a specific

Figure 6. Disagreements between two raters, Test A (left) and Test B (right).

script, a discrepancy had to be recorded. For example, for a script of Test A, if Rater A had awarded 55 marks and Rater B had awarded 40 marks, then a discrepancy was recorded and Rater A was judged to be more lenient on this occasion. The organization monitors the number of registered discrepancies per rater regularly.

For the purposes of this study, across the 10 scripts, if Rater A had appeared to be more lenient than Rater B more than once, then a directed link $A \rightarrow B$ was recorded in a rectangular matrix of dimensions 59×59 such that $y_{ij} = 1$, otherwise $y_{ij} = 0$ (so, we constructed a binary, directed network). Also, for the same 10 scripts, if Rater B was more lenient than Rater A more than once, then a directed link was recorded such that $y_{ji} = 1$, otherwise $y_{ji} = 0$. Apparently, it is possible for both $y_{ij} = 1$ and $y_{ji} = 1$, thus mutual links are possible in the network. The process resulted to a square matrix of $59 \times 58 = 3422$ unique pairs of comparisons. The process was repeated for both Tests A and B, so two datasets were constructed.

Fundamental Network Statistics

For Test A, the density was 37.5% (there were 1284 directed links out of the 3422 possible ones that could have been observed in the network). The raters had an average in-degree of 21.8 (see Figure 6, left panel). The correlation between in-degree and out-degree was -0.79 ($p < .001$), which is a positive health signal in the sense that the incoming and outgoing links should form competitively, as we would expect according to our theory (more incoming links demonstrate a more severe rater who is less likely to establish outgoing links). In total, there were $118/2 = 59$ mutual links in the data. Two of the raters demonstrated a very high number of mutual links. For example, the rater with ID = 6 had 14 mutual links with other raters, which is an indication of high inconsistency for that rater. In fact, four raters (7% of the 59 raters) were responsible for the 70% of the mutual links in the network. The ratings of these raters need to be scrutinized and the raters may need to receive more training or refrain from the operational rating.

Figure 7. A subgraph demonstrating the raters involved in cyclic triads (Test A). The size of the node is analogous to the number of total disagreements of the rater with the rest of the group. Squares indicate more experienced raters and circles indicate less experienced raters.

There was a very high % of weak transitivity triads for Test A; more specifically, 91% of all triad arrangements were of the weak transitivity kind, where $A \rightarrow B$, $B \rightarrow C \Rightarrow A \rightarrow C$. This is a positive health signal in the sense that we would expect—in a perfect world—all triads to be of this kind. On the other hand, three cyclic triads were observed; this needs to be investigated more closely to identify the raters involved (see Figure 7).

Figure 7 shows the raters involved in the three cyclic triads of Test A data. In the upper-right triangle of the figure, raters with IDs 22, 43, and 56 are involved in a cyclic triad. Rater 56 was more lenient than rater 43 (on at least 2 occasions out of 10 scripts). Rater 43 was more lenient than Rater 22 (again, on at least two occasions). However, Rater 22 appeared to be more lenient than Rater 56 (again, on at least two occasions); this is counterintuitive and an apparent source of inconsistency in the data. Rater 55 is involved in two cyclic triads and Rater 56 is involved in all three cyclic triads. As quantitative methods can take us that far, it would now be the time for practitioners to qualitatively eyeball the actual ratings of the three raters in order to pinpoint the source of the inconsistencies in the data. At the stage of exploratory investigations, one should not rule out misunderstandings, haphazard applications of the rubrics, or simply data entry mistakes.

The raters seem to be a little less well prepared to rate for Test B compared with Test A: The density for Test B was 53% compared with 37.5% for Test A (see Table 3). There was a larger number of mutual links ($N = 570/2 = 285$), compared

Table 3. Network Characteristics of the Rating Data of Tests A and B and Simulated Datasets.

	Test A	Test B	Mean of 1000 simulated random datasets	Mean of 30 simulated Rasch datasets
	using Test B characteristics			
Density (at least 2 significant discrepancies out of 10 scripts)	37.5%	53.0%	Set to 53%, as Test B	59.2% ($SD = 1.5$)
Reciprocity (total number of mutual links observed)	118/2 = 59	570/2 = 285	960/2 = 480 ($SD = 18.3$)	762/2 = 381 ($SD = 43.3$)
Cyclic triads (total number observed)	9/3 = 3	36/3 = 12	1005/3 = 335 ($SD = 49.2$)	30/3 = 10 ($SD = 4.9$)
% Weak transitivity ($i \rightarrow j$, $j \rightarrow k \rightarrow i \rightarrow k$)	91.0%	84.4%	53% ($SD = 0.9$)	88.6% ($SD = 0.87$)
% Strong transitivity ($i \rightarrow$ $k \rightarrow i \rightarrow j, j \rightarrow k$)	69.4%	62.3%	48.7% ($SD = 0.4$)	63.2% ($SD = 0.37$)
Correlation between in-degree and out-degree statistics	-0.79 ($p < .001$)	-0.82 ($p < .001$)	-0.02($SD = 0.12$)	-0.87 ($SD = 0.03$)

with Test A; three of the raters demonstrated a very high number of mutual links (they had 36, 31, and 25 mutual links, respectively). On average, there were 9 to 10 mutual links per rater and 12 cyclic triads were observed. This suggests a degree of randomness in the data and the raters involved need to be inspected in depth. However, the correlation between in-degrees and out-degrees was -0.82 ($p < .001$), which is a desirable characteristic of rating network data. Finally, 84.4% of all triads were weak transitivity triads. This is a lower proportion compared with Test A, but it is still high.

In the next two sections, we will demonstrate how we can investigate whether severity is stable across tests and whether there is a relationship between rater background characteristics and rater effects.

The Stability of Rater Characteristics Across Tests

There was a high correlation between the in-degree statistics of the raters for Tests A and B ($r = 0.63$, $p < .001$). There was also a high correlation between the out-degree statistics of the raters for Tests A and B ($r = 0.57$, $p < .001$). These are comparable to the correlations reported in the past using Rasch methodology in a similar context (Lamprianou, 2006). This suggests that raters who tend to be lenient (or severe) on one test will also tend to be lenient (or severe) on the other test. However, there is no correlation between the numbers of mutual links of the raters for the two tests; this is a suggestion that, although severity is a personal characteristic which remains roughly stable across tests, inconsistency may be less stable in the context of the specific qualifying sessions.

Investigation of the Relationship Between Rater Personal Characteristics and Rater Effects

The relationship between rater characteristics and rater effects was investigated using a directed ERGM. Two main effects were investigated for their role in the formation of in-degrees and out-degrees: (a) the years of teaching experience in public schools and (b) the rating experience of the raters. Teaching experience is a numeric variable that ranges from 8 to 35 years ($M = 23$ years, $SD = 5.3$). Thirty-five of the 59 raters had participated in the rating cycle of the previous year (coded as a binary variable where 1 means that the rater had participated in the previous year's rating cycle and 0 otherwise).

Table 4 illustrates the results of the same ERGM model for the data of Tests A and B. Only the main effects of the two background variables are modelled. The negative sign of the coefficient of the parameter "Mutual" suggests that any new edge has a very small likelihood to form a mutual link in the network (the coefficients are statistically significant for both Tests A and B). The negative coefficient for the "rater experience" variable suggests that those raters who had participated in the previous year's rating exercise had a smaller likelihood to form in- or out- edges, so

Table 4. Exponential Random Graph Model Investigating the Effect of Rater Characteristics on the Formation of In-degrees and Out-degrees.

	Test A		Test B	
	Coefficient	Standard error	Coefficient	Standard error
Edges	0.549	0.105	1.208	0.372
Mutual	−2.490	0.152	−2.018	0.114
Nodecov (teacher experience)	−0.027	0.008	−0.027	0.008
Nodecov (teacher experience)	0.018	0.008	0.046	0.008
Nodecov (rater experience)	−0.388	0.083	−0.528	0.082
Nodecov (rater experience)	−0.236	0.083	−0.172	0.082
AIC	4092.027		4236.773	
BIC	4128.855		4273.601	

Note. Nodecov represents the main effect of the variable in parenthesis on the likelihood of the formation of an out-degree. Nodecov represents the main effect of the variable in parenthesis on the likelihood of the formation of an in-degree. Mutual represents the likelihood a new edge to form a mutual link (provided the reciprocal edge already exists). AIC = Akaike information criterion; BIC, Bayesian information criterion.

rating experience seems to have a positive impact on the quality of ratings (the coefficients are statistically significant for both Tests A and B). On the other hand, teaching experience seems to decrease the likelihood that a rater will form incoming edges but increases the probability that a rater will form an outgoing edge.

There are two very popular techniques which are often used to evaluate the credibility of the results of an ERGM. First, the Monte Carlo Markov Chains estimation is investigated both graphically and through summary statistics. For example, we would typically wish to observe well-mixed, stationary chains, exploring thoroughly the parameter space. In addition, various goodness-of-fit procedures are often used to investigate model–data fit. Goodness of fit procedures simulate networks using the ERGM estimates and, for various network statistics, compare the distribution in the simulated networks to the observed values. For the sake of brevity, we will not delve into this further because it goes beyond the aims of this article. The interested reader could review Hunter, Handcock, Butts, Goodreau, and Morris (2008) for more details.

Comparison of the Rasch Analysis Output With the SNA/ERGM Output

A Rasch model was fit on the empirical data in order to investigate two fundamental psychometric characteristics of the raters, namely their severity (i.e., the Rasch rater severity estimate) and their consistency (the infit and outfit mean square).

For Test A, raters’ severity ranged from −0.44 logits to 0.22 logits ($M = 0$, $SD = 0.12$). In the context of the SNA approach, one could consider the in-degrees and/or

the out-degrees of a rater to be proxies of his or her severity. The in-degrees, for Test A, ranged from 2 to 58, with a mean of 21.7 ($SD = 13.7$). The out-degrees ranged from 0 to 56 ($SD = 18.1$). The correlation between raters' Rasch severity estimate and the in-degree statistic was 0.94 ($p < .001$). The correlation between raters' Rasch severity estimate and the out-degree statistic was -0.88 ($p < .001$).

For Test B, raters' severity ranged from -0.52 logits to 0.46 logits ($M = 0$, $SD = 0.15$). The in-degrees, for Test B, ranged from 0 to 58, with a mean of 30.7 ($SD = 15.2$). The out-degrees ranged from 2 to 58 ($SD = 16.1$). The correlation between raters' Rasch severity estimate and the in-degree statistic was 0.90 ($p < .001$). The correlation between raters' Rasch severity estimate and the out-degree statistic was -0.89 ($p < .001$).

As far as inconsistency is concerned, the correlation between the infit mean square statistic for the raters and their number of mutual links was $r = 0.75$ ($p < .001$) for Test A and $r = 0.76$ ($p < .001$) for Test B. The correlation between the outfit mean square statistics for the raters and their number of mutual links was smaller ($r = 0.39$ for Test A and $r = 0.58$ for Test B; $p < .001$ for both correlations). Overall, it makes sense to suggest that, in the context of the SNA approach, additional measures of inconsistency could be used. One, for example, could propose that a combination of the in-degree and out-degree measures can be a proxy of inconsistency. A rater who has a very high in-degree would normally not be expected to have a high out-degree because that would be counter-intuitive. Thus, a linear regression with infit mean square as the dependent variable and the interaction of in-degrees and out-degrees as independent variables gave an R^2 of .72 for Test A and .70 for Test B.

A Comparison Between Empirical Data and Random Simulated Data

We simulated 1000 random graphs with the same number of nodes, edges, and density as our empirical data of Test B (for the sake of brevity, we will not replicate the exercise for Test A because the inferences are comparable). The results are illustrated in Table 3. The mean number of mutual links was $960/2 = 480$ ($SD = 18.3$), the mean number of cyclic triads was $1005/3 = 335$ ($SD = 49.2$) and the mean correlation between in-degrees and out-degrees was -0.02 ($SD = 0.12$). Comparing the network characteristics of the random and the empirical data, one may observe that the latter have substantially smaller prevalence of reciprocity and cyclic triads and substantially higher prevalence of transitivity. It may be easily deduced that the data generation mechanism of our empirical data is far from being random.

A Comparison Between Empirical Data and Rasch Simulated Data

Comparing the empirical data with a random dataset may be a little bit far-fetched in the sense that the raters had received considerable training and many of them were experienced. It would be surprising, then, if the raters had produced their ratings randomly. However, it makes sense to compare the empirical data with simulated data

using one of the established psychometric models, such as the Rasch model. Simulated datasets using the Rasch model would provide a theory-based data-generation mechanism with which to compare our empirical data.

Hence, we used the Rasch model to simulate 30 datasets using the same number of raters, the same number of scripts and the same distribution of rater severity and script ability. First, we simulated the data, and then we analyzed each of the dataset using the proposed SNA approach. The results are illustrated in Table 3. The mean network density of the simulated data was 59.2% ($SD = 1.5\%$). The mean number of mutual links was 381 ($SD = 43.3$), the mean number of cyclic triads was 10 ($SD = 4.9$) and the mean correlation between in-degrees and out-degrees was -0.87 ($SD = 0.03$). The weak and strong transitivity indices were 88.6% ($SD = 0.87$) and 63.2% ($SD = 0.37$), respectively.

Directly comparing the empirical and the simulated data is not straightforward because for the generation of the simulated data we can only use the estimated parameters (the “true” rater severities and student abilities are not known). However, studying concurrently the network characteristics of the empirical, the random simulated and the Rasch simulated data we can safely deduce that the empirical data are much closer to the Rasch simulated data rather than to the random data.

Discussion

Human raters are routinely used in all sorts of high-stakes contexts, in order to evaluate students, professionals, products, programs, and so on. The academic community has reacted to this high demand, and has proposed a number of statistical methods that can be used to explore data from rating exercises. The most prominent of these methods are the Rasch model, generalizability theory and multilevel models. These methods are routinely used in order to evaluate the readiness of groups of raters to rate in a consistent manner; they are also used to identify individual raters who may need to be retrained or removed from operational rating.

Our research uses simulated and empirical data from two high-stakes assessments, to propose a new approach, based on SNA and ERGM, to evaluate the readiness of a group of raters for operational rating. Our proposal is motivated by the practicality and the intuitive nature of SNA visualizations; that means that SNA graphs can be a useful and practical tool to be used by policy makers, practitioners, and researchers who are less statistics-savvy. We propose and show how to use ERGM and Monte Carlo simulations (e.g., see Table 3), in order to draw inferences about the dataset network characteristics.

To summarize, in this article we introduced some basic concepts and some notation of the SNA and ERGM approaches and showed that both are appropriate and useful when analyzing rating data. We presented an instructive example of a simple SNA/ERGM analysis using simulated data and we coupled this with an analysis of empirical data. The results of this innovative method were compared to the results of

a Rasch analysis, which is a more “traditional” approach for the analysis of rating data. We demonstrated that there is a correspondence between specific Rasch and SNA/ERGM statistics. Finally, we also demonstrated how the proposed approach can be used to investigate the stability of rater characteristics and how the raters’ background characteristics relate to their rating behavior. These two practical examples are important because there is a vibrant community of researchers interested in conducting research with these two (and other similar) questions.

In our study, a distinction has been made between agreement and disagreement networks. The decision between choosing to analyze an agreement rather than a disagreement network can be heavily context-dependent. Most frequently, in well-established rating systems such as large-scale assessments, the raters are recruited from pools of experienced individuals. In these contexts, most raters will have high degrees of agreement with the other raters; as a consequence, agreement networks will be very busy, with too many edges, and will probably be more difficult to interpret than disagreement networks. On the other hand, in early stages of rater training, and especially when new rating rubrics are being developed, agreement networks may be more useful because they will allow the easier identification of clusters of raters who “think alike.” One may acknowledge, however, that experience and practice will eventually suggest what is most appropriate in each case.

Because of the vast volume of the literatures involved, it was not possible to cover everything in this article. More research is certainly needed in order to investigate—possibly using qualitative methods as well—how practitioners can adapt the proposed SNA/ERGM approach in their own particular contexts. Also, more research is needed to investigate how the proposed approach can be used in order to prepare comprehensive rater profiling systems that can provide formative and personalized feedback to individual raters before operational rating commences. It is also necessary to investigate how more informative networks, for example, valued networks, can be practically adapted for use in the context of the analysis of rater data.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material

Supplementary material for this article is available online.

Note

1. The interested reader may also refer to Garner and Engelhard (2009) for Rasch estimation algorithms using relevant techniques.

References

- Baird, J., Hayes, M., Johnson, R., Johnson, S., & Lamprianou, I. (2013). *Marker effects and examination reliability: A comparative exploration from the perspectives of generalizability theory, Rasch modelling and multilevel modelling*. Report to The Office of Qualifications and Examinations Regulation (Ofqual), UK.
- Banerji, M. (1999). Validation of scores/measures from a K-2 developmental assessment in mathematics. *Educational and Psychological Measurement*, 59, 694-715.
- Barrett, S. (2001). The impact of training on rater variability. *International Education Journal*, 2, 49-58.
- BBC. (2003, February 25). *New drive for exam markers*. Retrieved from http://news.bbc.co.uk/go/pr/ft/-/1/hi/northern_ireland/2797651.stm
- Block, P. (2015). Reciprocity, transitivity, and the mysterious three-cycle. *Social Networks*, 40, 163-173. doi:10.1016/j.socnet.2014.10.005
- Bojanowski, M., & Corten, R. (2014). Measuring segregation in social networks. *Social Networks*, 39, 14-32. doi:10.1016/j.socnet.2014.04.001
- Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, 20, 89-110.
- Braun, H. I. (1988). Understanding score reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics*, 13, 1-18.
- Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer-Verlag.
- Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, 37, 163-178.
- Cushing, S. W. (1998). Using FACETS to model rater training effects. *Language Testing*, 15, 263-287.
- Cushing, S. W. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, 6, 145-178.
- Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: a many-facet Rasch analysis. *Language Assessment Quarterly*, 2, 197-221.
- Engelhard, G., Jr. (1992). The measurement of writing competence with a many-faceted Rasch model. *Applied Measurement in Education*, 5, 171-191.
- Engelhard, G., Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, 31, 93-112.
- Engelhard, G., Jr. (1997). Constructing rater and task banks for performance assessments. *Journal of Outcome Measurement*, 1, 19-33.
- Engelhard, G., Jr. (2002). Monitoring raters in performance assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large scale assessments for all students: Validity, technical adequacy, and implementation* (pp. 261-288). Mahwah, NJ: Lawrence Erlbaum.
- Engelhard, G., Jr. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York, NY: Routledge.
- Freedman, S. W. (1981). Influences on evaluators of expository essays: beyond the text. *Research in the Teaching of English*, 15, 245-255.

- Foster, J. G., Foster, D. V., Grassberger, P., & Paczuski, M. (2010). Edge direction and the structure of networks. *Proceedings of the National Academy of Sciences of the U S A*, 107, 10815-10820.
- Garner, M., & Engelhard, G. (2009). Using paired comparison matrices to estimate parameters of the partial credit Rasch measurement model for rater-mediated assessments. *Journal of Applied Measurement*, 10, 30-41.
- Goldstein, H. (2010). *Multilevel statistical models* (4th ed.). London, England: Arnold.
- Goodreau, S. (2007). Advances in exponential random graph (p^*) models applied to a large social network. *Social Networks*, 29, 231-248.
- Harris, J. K. (2014). *An introduction to exponential random graph modeling*. Thousand Oaks, CA: Sage.
- He, Q., Anwyll, S., Glanville, M., & Deavall, A. (2013). An investigation of the reliability of marking of the Key Stage 2 National Curriculum English writing tests in England. *Educational Research*, 55, 393-410.
- Hoskens, M., & Wilson, M. (2001). Real-time feedback on rater drift in constructed-response items: An example from the Golden State Examination. *Journal of Educational Measurement*, 38, 121-145.
- Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M., & Morris, M. (2008). ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software*, 24(3), 1-29.
- Lamprianou, I. (2004). *Marking quality assurance procedures for large-scale high-stakes assessment* (Not in public domain), National Assessment Agency, UK.
- Lamprianou, I. (2006). The stability of marker characteristics across tests of the same subject and across subjects. *Journal of Applied Measurement*, 7, 192-200.
- Lamprianou, I. (2008). High-stakes tests with self-selected essay questions: Addressing issues of fairness. *International Journal of Testing*, 18, 55-89.
- Lamprianou, I., & Boyle, B. (2004). Accuracy of measurement in the context of mathematics National Curriculum tests in England for ethnic minority pupils and pupils who speak English as an additional language. *Journal of Educational Measurement*, 41, 239-260.
- Leckie, G., & Baird, J. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement*, 48, 399-418.
- Linacre, J. M. (1994). *Many-facet Rasch measurement* (2nd ed.). Chicago, IL: MESA Press.
- Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, 3, 331-345.
- Lunz, M. E., Stahl, J. A., & Wright, B. D. (1991). *The invariance of judge severity calibrations*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- Masters, G. N. (2002). *Fair and meaningful measures? A review of examination procedures in the NSW Higher School Certificate*. Camberwell, Victoria, Australia: ACER Press.
- McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27, 415-444.
- Myford, C. M., & Wolfe, E. W. (2002). When raters disagree, then what: Examining a third-rating discrepancy resolution procedure and its utility for identifying unusual patterns of ratings. *Journal of Applied Measurement*, 3, 300-324.

- Rae, G., & Hyland, P. (2001). Generalizability and classical test theory analyses of Koppitz's Scoring System for human figure drawings. *British Journal of Educational Psychology*, 71, 369-382.
- Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded edition, 1980. Chicago, IL: University of Chicago Press)
- Robins, G., Pattison, P., Kalish, Y., & Lusher, D. (2007). An introduction to exponential random graph (p*) models for social networks. *Social Networks*, 29, 173-191.
- Royal-Dawson, L., & Baird, J.-A. (2009). Is teaching experience necessary for reliable scoring of extended English questions? *Educational Measurement: Issues and Practice*, 28(2) 2-8.
- Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, 88, 413-428.
- Smith, R. M. (1991). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement*, 51, 541-565.
- Smith, R. M. (2000). Fit analysis in latent trait measurement models. *Journal of Applied Measurement*, 1, 199-218.
- Smithers, R. (2003, January 31). Exams chief warns of more chaos this summer. *The Guardian*. Retrieved from <https://www.theguardian.com/uk/2003/jan/31/politics.alevels2002>
- Snijders, T. A., Van de Bunt, G. G., & Steglich, C. E. (2010). Introduction to stochastic actor-based models for network dynamics. *Social Networks*, 32, 44-60.
- Stahl, J. A., & Lunz, M. E. (1991). *Judge performance reports: Media and message*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.
- Wang, J., Engelhard, G., Jr., & Wolfe, E. W. (2016). Evaluating rater accuracy in rater-mediated assessments using an unfolding model. *Educational and Psychological Measurement*, 76, 1005-1025. doi:10.1177/0013164415621606
- Wang, W.-C., & Chen, C.-T. (2005). Item parameter recovery, standard error estimates, and fit statistics of the Winsteps program for the family of Rasch models. *Educational and Psychological Measurement*, 65, 376-404. doi:10.1177/0013164404268673
- Weigle, S. (1998). Using FACETS to model rater training effects. *Language Testing*, 15, 263-287.
- Williams, O., & Del Genio, C. I. (2014). Degree correlations in directed scale-free networks. *PLoS One*, 9:e110121.
- Wiseman, C. S. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing*, 17, 150-173.
- Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.
- Wright, B. D., & Mok, M. (2000). Rasch models overview. *Journal of Applied Measurement*, 1, 83-106.
- Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA Press.
- Wu, S. M. & Tan, S. (2016). Managing rater effects through the use of FACETS analysis: The case of a university placement test. *Higher Education Research & Development*, 35, 380-394. doi:10.1080/07294360.2015.1087381