ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(4); 492-507

An Illustration of a Latent Class Analysis for Interrater Agreement: Identifying Subpopulations with Different Agreement Levels

Ömer Emre Can ALAGÖZ* Yılmaz Orhun GÜRLÜK** Mediha KORKMAZ***
Gizem CÖMERT****

Abstract

This study illustrates a latent class analysis (LCA) approach to investigate interrater agreement based on rating patterns. LCA identifies which subjects are rated similarly or differently by raters, providing a new perspective for investigating agreement. Using an empirical dataset of parents and teachers evaluating pupils, the study found two latent classes of respondents, one belonging to a moderate agreement pattern and the other belonging to low agreement pattern. We calculated raw agreement coefficient (RAC) per behaviour in the whole sample and each latent class. When RAC was calculated in the whole sample, many behaviour had low/moderate RAC values. However, LCA showed that these items had higher RAC values in the high agreement and lower RAC values in the low agreement class.

Keywords: Interrater Agreement, Latent Class Analysis, Raw Agreement Coefficient, Agreement Methods, Mixture Modelling

Introduction

Using self-report methods for measuring unobservable psychological constructs (i.e., latent variables, traits, factors) is sometimes not possible due to several reasons. For example, researchers need external raters (i.e., observers, evaluators) to gather information about the subjects when the study focuses on disadvantaged groups, students or infants. More specifically, asking children to describe their aggression level by means of filling a questionnaire might be an unrealistic goal. Rather, researchers may employ teachers as raters to assess students' aggression level by observation. Generally, researchers prepare an evaluation list or a manual to inform raters about the indicators of the latent constructs and how to rate them. These ratings can be based on a behaviour's presence/absence, traits, or frequency of behaviour (Bıkmaz Bilgen & Doğan, 2017; Tanner & Young, 1985; Uebersax, 1990; Von Eye & Mun, 2005).

These rating scores are used for different purposes such as research (Leising et al., 2013) or diagnosis of mental illness (Shaffer et al., 1993). Therefore, it is very important to give objective and reliable information to the raters, otherwise, any result from their ratings is likely to be inaccurate. To improve the rating accuracy, researchers usually employ multiple raters. These raters are expected to rate a subject in a similar way since they all follow the same objective instructions. Therefore, using multiple raters and evaluating the consistency between them can also help researchers to understand the quality of instructions (e.g., clarity, objectivity). This consensus among the raters is referred to as interrater agreement (Hallgren, 2012; Landis & Koch, 1977; Uebersax, 1990). Researchers have suggested several methods for testing interrater agreement. These methods can be collected under two headings: 1) classical methods 2) latent variable methods. Since this study focuses on discrete variables, we are only

* Research Assistant, University of Mannheim, alagoez@uni-mannheim.de, ORCID ID: 0000-0003-3305-6564
** PhD student, Ege University, Faculty of Literature, İzmir-Türkiye, yilmazorhungurluk@gmail.com, ORCID ID: 0000-0002-1134-3776
*** Assoc. Prof., Ege University, Faculty of Literature, İzmir-Türkiye, medihakrkmz@gmail.com, ORCID ID: 0000-0001-6504-5822
****PhD student, Ege University, Faculty of Literature, İzmir-Türkiye, cmrtgizem@gmail.com, ORCID ID:0000-0001-7555-6378

To cite this article:
Alagöz, Ö.,E.,C., Gürlük, Y.,O., Korkmaz, M. & Cömert, G. (2023). An illustration of a latent class analysis for interrater agreement: identifying subpopulations with different agreement levels. *Journal of Measurement and Evaluation in Education and Psychology, 14*(4), 492-507. https://doi.org/10.21031/epod.1308732

Received: 1.06.2023
Accepted: 23.11.2023

concerned with the methods suitable for them. Belonging to classical methods, Cohen's Kappa (Cohen, 1960), Fleiss' Kappa (Fleiss, 1971), and Krippendorff's Alpha coefficients are three of the most popular interrater agreement tests for discrete variables (see footnote 1). However, there is a disadvantage to using them, namely these coefficients are biased when most of the raters only use a specific rating category (Gisev et al., 2013; Göktaş & İşci, 2011, Hayes & Krippendorff, 2007; Yarnold, 2016). In such cases, it is suggested that bias caused by the sparse contingency matrix can be prevented by using raw agreement coefficient (RAC; i.e., percentage of agreement; Feinstein & Cicchetti, 1990; Viera & Garrett, 2005).

In recent years, interrater agreement is studied with latent variable-based probability statistics. Although approaches based on probability distributions were widespread during the 60s, usage of these methods has become more common thanks to software developments (Jiang, 2019; Raykov et al., 2013). There are several advantages to using latent variable models for measuring the interrater agreement. First, we can test the rating consistency between different raters (Agresti, 1992; Yilmaz & Saracbasi, 2017; Yilmaz & Saracbasi, 2019). Second, we can extract the agreement patterns. Third, we can detect anomalies in these patterns. Fourth, we can calculate the sensitivity of raters who give the same rating to participants. Finally, by comparing the agreement patterns obtained from different measurements of the same construct (e.g., two independent studies using the same test with multiple raters), latent variable approaches inform researchers about the reliability and validity of the measurement tool (Jiang, 2019; Kottner et al., 2011; Tanner & Young, 1985; Schuster & Smith, 2002).

Here, we briefly explain two previous latent variable approach to interrater agreement, one treating agreement continuous and one treating it discrete. The first approach is a confirmatory factor analysis that is proposed by Raykov and colleagues (2013). In their approach, rating pattern of a rater can be examined with category thresholds. It is important to note that these category thresholds are rater specific, which means that we do not obtain a summary rating pattern for the whole sample, but we obtain rating pattern for each rater separately. One can test the identity of thresholds across raters to investigate the invariance of cut-offs that raters use to evaluate subjects. By examining the rater-specific thresholds, one can identify those raters who have aberrant rating pattern.

The second latent variable approach is a Latent Class Analysis (LCA) model, which is proposed by Schuster and Smith (2002). Their LCA model consists of two categorical latent variables. The first latent variable represents the true class of a subject, and the second latent variable represents whether a subject is obvious (i.e., all raters agree on them) or ambiguous (i.e., at least one rater disagrees with the rest). Raters can easily identify the true class membership of an obvious subject, whereas they randomly guess the class membership of an ambiguous subject. The agreement on the obvious and ambiguous subjects are referred to as *systematic agreement* and *chance agreement*, respectively. The former can be interpreted as the true agreement between raters. Note that LCA can be used when the rating is done with categorical variables. If rating is done with continuous variables, one can use another approach that is based on Latent Profile Analysis (LPA; Major et al., 2018). For some other LCA approaches in interrater agreement studies, we refer readers to Basten and colleagues (2015), De Los Reyes and colleagues (2009), and Major and colleagues (2018).

To our knowledge, classical approach or latent variable modelling, interrater agreement methods and studies using them focus merely on the consistency between raters (e.g., Forster et al., 2007; Miller, 2011; Thomson, 2003). Indeed, the consistency between raters are vital parts of studies using multiple raters, but another important question arises whenever there is no perfect agreement between raters: Do raters disagree on every subject for every item? In other words, we are interested in whether there are different sets of subjects, each of which is rated in a different way by the raters. There could be, for instance, one set of subjects that are similarly rated by raters for most of the items, and another set of subjects that are rated differently by raters for most of the items. If such a heterogeneity is ignored and an interrater agreement method is used, the result is likely to show a disagreement. In such a case, an important piece of information is disregarded: the subset of subjects on whom raters agreed. Therefore, we propose a LCA approach that detects latent classes of subjects which differ in how they are rated. In the method section, we describe the details of the proposed LCA approach. Then, we analyse an empirical data set and interpret the results.

Sample

In the study, some data within the framework of the "I'm Learning to Protect Myself with Mika" sexual abuse prevention program developed by Kızıltepe and colleagues (2022) were used with the permission of the researchers (see footnote 2). In order to determine whether the aforementioned intervention program had side effects, the researchers created 10 dichotomously scored items (for item contents, see Table 3). The parents evaluated only their children and the teachers evaluated all of the students who have taken the MIKA prevention program. For the examined agreement between teachers and parents two rater blocks were structured as parent and teacher. The blocks were handled as two raters, and it was tested whether there was concordance between the evaluations of the teachers and the parents.

The sample of the study consists of 290 children in the 5-year-old age group from the lower, middle, and upper socio-economic status attending kindergarten and their parents. In the study, considering the districts where the kindergartens are located, two schools from the lower socio-economic status, three from the middle socio-economic status and one school from the upper socio-economic status were selected. Two of the schools are private and four of them were kindergartens within the part of a public institution. This form was answered by parents and teachers for 290 children. The proportion of boys and girls in the sample were 52.76% (*N*=153) and 47.24% (*N*=137), respectively. After data screening, 13 observations were omitted (Both parents and teachers of 4 and only parents of 9 did not evaluate the children). In order to create intervention and control groups that are equivalent in terms of age, socio-economic status, gender and so on, one classroom from each school was included in the training group and one classroom in the control group. Only the intervention group was used to examine the agreement between raters. The ages of the children ranged from 50-72 months (Mean = 61.80, SD = 6.1). The ages of the mothers of the children participating in the study ranged from 22 to 47 (Mean = 24.32, SD = 5.25), while the age of their fathers ranged between 25 and 53 (Mean = 37.67, SD= 5.59). In the study, 84% of the parents participating were mothers and 16% were fathers.

Methods

LCA Approach to Interrater Agreement

We utilize LCA to detect subpopulations of subjects that differ by how they are rated on several categorical items by different raters. Different from other LCA rater agreement approaches, our approach does not classify raters into "agreement" and "disagreement" classes and does not investigate the similarity of ratings at item-level. Rather, we classify subjects into classes depending on the similarity of scores given to them by different raters. Therefore, this approach does not necessarily find "agreement" or "disagreement" classes, but it captures whether raters evaluate all respondents similarly on a set of items.

First, we transform the data set into a new form by subtracting the scores given by one group of raters from the scores given by the other group of raters. That is, assume that X^A and X^B are $N \times J$ matrices of scores given to N number of subjects on J number of items by rater group A and rater group B, respectively. Then, the matrix of rating distances between rater groups, X^{A-B}, is an $N \times J$ matrix that is calculated by $X^A - X^B$. If raters evaluate subjects on $M \in \{1, \dots, m\}$ response categories, $X_{nj}^{A-B} \in \{1 - m, \dots, 0, \dots, m-1\}$ is the signed distance between two raters for subject *n* and variable *j*. A negative (positive) X_{nj}^{A-B} means that the subject is assigned a higher (lower) score by the rater in group B than the rater in group A. If X_{nj}^{A-B} is zero, then subject *n* is assigned the same score by raters in both groups, therefore raters agree with each other.

Second, we conduct LCA with the transformed data set X^{A-B}. Each column of X^{A-B} is specified as an indicator variable. Therefore, classes are defined with how divergent do raters score subjects. We fit models with increasing number of latent classes and investigate several fit indices to decide on the number of latent classes. By investigating conditional response probabilities, we can find which subjects on which items are evaluated similarly (or differently) by the raters.

Third, we calculate the posterior class probabilities for each subject and assign them to the class for which their posterior class probability is the highest (i.e., modal assignment, Dias & Vermunt, 2008).

Optionally, if we believe classes represent subjects that are rated similarly and differently by raters, then we can conduct a classical rater agreement analysis (e.g., Kappa, RAC) in each class to verify whether it was the case.

Latent classes can differ due to presence/absence of agreement, type of disagreement, and items where these (dis-)agreements occur. If latent classes differ due to presence/absence of agreement, then the response probability in one class will be the highest for the category "0", whereas in other classes it will be the lowest for the category "0". If classes differ regarding the type of disagreement (see footnote 3), response probabilities in one class(es) can be higher for positive categories, whereas in the other class(es), it can higher for negative categories. Finally, the reason for class differences can differ item by item. That is for one item, classes can differ due to presence/absence of agreement, and for another item, they can only differ due to type of disagreement. In conclusion, it is very important to carefully examine the conditional response probabilities to make sense of classes. Indeed, it is a rule that applies to all latent class analyses.

Procedure

In the first step, we conducted LCA in the abovementioned way. Then, the number of classes that differ in the similarity of ratings between two rater groups were determined, and respondents were assigned into the class for which their posterior class probability was the highest. Next, we investigated the conditional response probabilities for rating distances to understand the characteristics of each class. Finally, we calculated the RAC in each class to confirm our interpretations about class characteristics. We used Latent GOLD (Vermunt & Magidson, 2008) for conducting LCA and IBM SPSS Statistics 25 for calculating RAC. In our empirical illustration, we calculated RAC due to skewed ratings provided by one rater group. Hereby we explain how RAC can be calculated. Given that two raters evaluate *N* number of subjects on *M* number of categories, the below $M \times M$ table can be constructed. There, n_{jk} denotes the total number of subjects that are rated with category *j* by rater A and with category *k* by rater B. On diagonal, we have numbers of cases where raters evaluate subjects with the same categories.

Table 1:
Summary of ratings by two raters. $M \times M$ table of ratings provided by two raters, where each cell n_{jk} represents the number of subjects who are rated with category j by rater A and with category k by rater B.

	Rater A				
Rater B	1	2	…	M	Column Sums
1	n_{11}	n_{12}	…	n_{1M}	n_{1+}
2	n_{21}	n_{22}	…	n_{2M}	n_{2+}
…	…	…	…	…	…
M	n_{M1}	n_{M2}	…	n_{MM}	n_{M+}
Row Sums	n_{+1}	n_{+2}	…	n_{+M}	N

Then RAC is calculated as the proportion of respondents rated with the same category by both rater A and rater B:

$$RAC = \frac{1}{N}\sum_{m=1}^{M} n_{mm}$$

Results

In line with the model described in the method section, we fit latent class models with 1 to 5 classes. To decide on the number of latent classes, we compared Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

AIC and BIC are information criteria that takes model fit and model complexity into account and inform us about the balance between fit and complexity. The lower the AIC and BIC get, there is a better balance between fit and complexity. Since including more parameters to the model will increase the model fit as it approaches to saturation, these criteria penalize the fit index with the number of observed units (i.e., sample size) in AIC and with both the number of observed units and number of parameters in BIC. For this reason, BIC is mostly preferred over AIC. Also for LCA, the model with the lowest AIC or BIC should be preferred, but in case of very close values, researchers can choose among best fitting models according to theory and interpretability of results (Nylund and colleagues, 2007).

In Table 2, we provide the AIC and BIC values for LCA model with different number of classes. Accordingly, as is seen in Table 2, AIC favours the model with 5 classes, whereas BIC suggests the model with 3 classes fit the data best. We base our decision on BIC results since it penalizes the model with more parameters. However, since the difference in BIC between the model with 2 classes and 3 classes is very small, we chose the model with 2 classes for parsimony and interpretability reasons. Also, a further investigation of the model with 3 classes showed that the size of the added class is very close to zero, which then made sense to choose the model with 2 classes. We care for brevity and parsimony because the reason for using LCA is to capture different rating patterns but not to make substantial inferences about classes. In this 2-class model, the size of Class 1 was found 0.62 (*N*=183) and of Class 2 was found 0.38 (*N*=107).

Table 2:

Model comparison. From the leftmost column to the rightmost column, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and degrees of freedom (df)

Number of Classes	AIC	BIC	Df
1	7171.56	7432.12	219
2	7021.89	7322.82	208
3	6977.21	7318.51	197
4	6972.28	7353.95	186
5	6957.54	7379.58	175

We investigated conditional probabilities (see Table 3) of the differences between parents' ($X^{parents}$) and teachers'($X^{teachers}$) ratings to understand in what sense these classes differ from one another. The most visible difference between classes is the probability of category zero, in other words, agreement. We see that subjects in Class 1 have higher probabilities of being identically rated by their parents and teachers on all items. Actually, the probability of agreement is generally very high and larger than .50 in all items except items 4, 5, and 6. However, in Class 2, the probability of category zero is roughly below .50 in all items except item 3.

In Table 3, we also see for Class 1 that the second largest category probability after "0" is "+1", and "+1" is followed by "-1" (item 6 can be an exception). Meaning of this pattern is, in Class 1, the most probable outcomes are either a perfect agreement or a difference of one unit between raters. This finding further supports that Class 1 is associated with high agreement. If we look at Class 2, we see that it is either the extreme disagreement category "+4" or intermediate disagreement categories "+1, +2, +3" have the highest probabilities (item 3 can be an exception).

The implication of this pattern is that Class 2 is associated with higher disagreement, where teachers provide low ratings and parents provide high ratings for the subjects.

The reason for disagreement between parents and teachers can be that teachers systematically evaluate their students lower than parents or parents systematically evaluate their children higher than teachers.

To investigate if either one is the case, we investigated the category frequencies and variances for parents and teachers. The typical finding was parents showed a higher variance in their ratings compared to teachers and teachers used smaller categories more often than parents. For example, the variances of parents' and teachers' ratings about "Anger Problems" item were 1.29 and 0.45, respectively. Moreover, parents rated 105 children with the lowest category, whereas teachers rated 254 students with the lowest category. Therefore, the disagreement occurs because teachers systematically give a low score to the children. The reason for such tendency could be that teachers have only limited time and a fixed context to observe children, whereas parents spend more time and observe their children in different contexts (also they have biases from the times before the study).

Table 3:

Conditional probabilities of rating differences between different raters $Pr\left(X_{nj}^{A-B} = m \middle| Class = c\right)$

Items	Classes	$m\ (X^{parents} - X^{teachers})$									RAC
		-4	-3	-2	-1	**0**	1	2	3	4	
Item 1	Class 1	-	-	.01	.02	**.71**	.23	.03	.01	.01	**.72**
	Class 2	-	-	.01	.01	**.37**	.38	.17	.04	.04	**.32**
Item 2	Class 1	-	.01	-	.01	**.71**	.20	.04	.01	.01	**.72**
	Class 2	-	.01	-	.01	**.46**	.26	.10	.06	.11	**.44**
Item 3	Class 1	-	-	.01	.01	**.82**	.12	.03	.01	-	**.84**
	Class 2	-	-	.01	.01	**.75**	.17	.06	.01	-	**.75**
Item 4	Class 1	.02	.07	.10	.15	**.28**	.20	.09	.03	.03	**.31**
	Class 2	.01	.03	.05	.10	**.26**	.24	.15	.07	.07	**.21**
Item 5	Class 1	.01	.03	.04	.10	**.47**	.25	.06	.02	.01	**.52**
	Class 2	.01	.01	.01	.03	**.28**	.31	.16	.10	.10	**.23**
Item 6	Class 1	-	.03	.03	.06	**.24**	.32	.18	.08	.06	**.25**
	Class 2	-	.01	.01	.01	**.07**	.18	.21	.20	.32	**.06**
Item 7	Class 1	-	.01	.01	.08	**.72**	.15	.03	.01	.01	**.69**
	Class 2	-	.01	.01	.02	**.52**	.27	.12	.04	.02	**.54**
Item 8	Class 1	.01	.02	.01	.05	**.62**	.21	.05	.01	.01	**.67**
	Class 2	.01	.01	.01	.01	**.31**	.28	.18	.09	.12	**.27**
Item 9	Class 1	-	.01	.01	.02	**.83**	.12	.01	.01	.01	**.86**
	Class 2	-	.01	.01	.01	**.51**	.27	.09	.07	.04	**.46**
Item 10	Class 1	.01	.01	.01	.09	**.69**	.18	.01	-	.01	**.71**
	Class 2	.01	.01	.01	.02	**.48**	.39	.05	-	.06	**.45**

Note. The cells with "-" means that the difference between parents' and teachers' was never yielded the category *m* for any respondent. Conditional probabilities of the agreement category "0" and class-specific raw agreement coefficient (RAC) are given in bold face.

Since there is a visible difference in the agreement probabilities between classes, we further investigated whether Class 1 that has higher agreement probabilities yielded a higher agreement coefficient with a classical interrater agreement analysis. Therefore, we calculated RAC for each item once for respondents in Class 1 and once for respondents in Class 2. When RAC in Class 1 was examined, we see values larger than .70 for five items and larger than .50 for eight items (see Table 4). These eight items are the ones that conditional response probabilities suggested similarity in ratings between parents and teachers. Furthermore, both conditional probabilities of "0" from LCA and small RAC values pointed out that there is a difference between parents' and teachers' rating patterns for items 5 and 6 in Class 1. For the results of Class 2, it was visible that all items have RAC smaller than around .50 except for item 3 and 7, for which conditional response probabilities also suggested dissimilar rating patterns between parents and teachers.

What is interesting is that conditional response probability of category "0" is almost identical to the class-specific RAC for all items. However, this is not much surprising since RAC is quantified by the total proportion of subjects that are evaluated with the same category by two raters, subtraction of which is equal to "0". Yet, RAC is merely calculated with the observed variables and deterministic, whereas the RAC-like values obtained via LCA is probabilistic. This finding implies that conducting LCA can be adequate to also quantify the interrater agreement without further separate analysis.

Table 4:

Raw Agreement Coefficients per item for Class 1 (left), for Class 2 (middle), for the whole sample (right).

Items	Class 1	Class 2	Overall
Afraid of Animals	0.72	0.32	0.57
Separation Anxiety	0.72	0.44	0.61
Questions about Sexuality	0.84	0.75	0.81
Difficulty of Expressing Emotions	0.31	0.21	0.27
Disobedience	0.52	0.23	0.41
Whining	0.25	0.06	0.17
School Avoidance	0.69	0.54	0.63
Anger Problems	0.67	0.27	0.52
Stranger Anxiety	0.86	0.46	0.71
Afraid of Adults	0.71	0.45	0.61
Mean	0.63	0.37	0.53

Discussion and Conclusion

In this study, we proposed a LCA to investigate rater agreement for categorical rating data. We first explained the need for and benefits of such an approach then described how LCA parameters help investigating rater agreement with a fictitious example. Lastly, we analysed an empirical dataset with the proposed approach.

Previous classical or latent variable approaches for rater agreement research focus on quantifying the agreement between two or more raters for all respondents at once. However, it is also possible that raters give similar scores to one set of respondents and different scores to another set of respondents. To detect such respondent groups, one can first create new variables by subtracting the ratings of one rater group from the ratings of another rater group. Hence, this new variable indicates the distance and its direction between the ratings of two rater groups. Then, one can conduct LCA on these new variables to capture respondent subpopulations who were rated similarly (zero distance) and differently (non-zero distances) by rater groups.

The current practice is to ignore the existence of such subpopulations associated with different levels of agreement and to calculate a single agreement coefficient (e.g., RAC) for the whole sample. However, in the presence of subpopulations, the sample RAC is roughly the weighted average of RAC values calculated for subpopulations. Hence, the sample RAC is likely to be biased. In our empirical example, we identified two latent classes, one related with smaller rating distances and the other related with larger rating distances between raters. Indeed, we showed that the RAC calculated for the whole sample was around the weighted average of class-specific RAC values. Furthermore, we showed that RAC calculated in smaller distance class was higher than the RAC calculated in larger distance class.

Another advantage of using LCA is that, in case of disagreement between raters, conditional response probabilities inform us about the direction of disagreement. If researchers conduct a classical agreement analysis, they obtain a single coefficient value quantifying the agreement. To understand more about why disagreement occurs, they need to examine descriptive statistics. However, the conditional response probabilities in LCA already tells researchers about which rater group tends to give higher or smaller scores than the other group. Moreover, researchers can also easily evaluate how severe is the disagreement. For example, disagreement is less severe when the conditional response probabilities are higher for ± 1 distance categories compared to when they are higher for ± 5 distance categories.

The reason for using RAC in our analysis was the sparse contingency tables of ratings. That is, some raters did not use some of the categories. In case of a sparse contingency table, other classical agreement analyses than RAC were found to yield biased estimates. As the second step after finding latent classes, one can always use other analyses within each class to see if they indeed represent different levels of agreement. However, calculating RAC for a variable after LCA was redundant in our analysis since the conditional response probability of distance category "0" has always corresponded to the RAC value of that specific class (see Table 2).

In the proposed LCA approach, we include all items to the analysis at once, whereas one calculates the agreement item by item in the classical rater agreement approaches. By doing so, all items contribute to the classification of respondents into classes. One can of course include only one set of items to the analysis, but it rather conflicts with the main idea of using LCA for rater agreement analysis, that is to make use of rating patterns across many items rather than doing item by item analyses.

As in all studies, there are also limitations to our approach. First limitation is the sample size requirement of LCA. Usually, it is required to have at least 500 respondents in the data set for using LCA. Although this requirement is not a limitation to the LCA approach to the rater agreement, it is for our empirical example. However, we do not see our sample size (i.e., 290 children) as a problem for two reasons: 1) the main aim for using LCA is to capture differences between rating patterns, and 2) the empirical example is only used to demonstrate how LCA parameters are interpreted in rater agreement domain but not to answer substantive research questions about the MIKA measurement tool.

Another limitation that we are aware of is the ambiguity of interpretations of classes. Actually, it is not a limitation but a general feature of LCA. That is, researchers need to examine class-specific parameters

to understand or speculate about the characteristics or the labels of classes. We see it as a limitation in the sense that our approach does not yield two clear cut classes related to rater agreement and disagreement. Indeed, in the empirical illustration, there were some items that contradict with our class explanations. However, we believe that the majority of items were consistent with our class interpretation. Yet, we acknowledged that such interpretations can be sometimes subjective, that is why we included Table 3 with detailed conditional probabilities to be transparent and to allow readers to better evaluate our interpretation (as ours is maybe only one out of many alternatives). Moreover, in Appendix B, we provide the results from the model with three classes, which was the favourite of BIC, to see if it leads to a different class explanation (see footnote 4). However, the newly added class was almost practically empty (N=9), so none of our interpretations have differed. Compared to our empirical example, some data sets might require more time and effort to make sense of the meaning of latent classes.

Future research can include respondent level covariates that might explain why they were assigned into different classes. If classes separate who are rated similar or different (as in our empirical analysis), then such covariates would tell why raters agreed or disagreed on the rating of an item. Also, future research can focus on a confirmatory mixture model that classifies raters into agreement and disagreement classes. Further, it would be interesting to see how this approach works with other data sets, both from similar and different domains, and whether similar types of classes arise with other data sets. Finally, future research should examine using LCA with non-sparse data to see whether other classical agreement analyses are also suitable for quantifying class-specific rater agreements in the second step.

Despite the limitations, we believe that latent variable models can help us learn more about the rater agreement. We also believe that our approach focuses on an important aspect of the rater agreement, which is the respondents that are being rated. With the proposed LCA approach, one does not have to disregard the substantive research question at hand because of the rater disagreement, but they can focus on the respondents for whom there is an agreement between raters.

Footnotes:

1) For interested readers, among many other, Konstantinidis and colleagues (2022), Zapf and colleagues (2016), Sertdemir and colleagues (2013), and Ato and colleagues (2011) provide extensive simulation studies and theoretical discussions regarding the classical agreement methods and the comparison of their performances. These methods and their comparisons are beyond the scope of this study, therefore they are not discussed in this paper.
2) Requests for accessing the data that support the findings of this study should be made to the the corresponding author of Kızıltepe and colleagues (2022).
3) It can co-exist with the presence/absence classes. For example, there can be one class with respondents who are scored similarly, another class with respondents for whom raters in group A provide higher scores than raters in group B, and another class with respondents for whom raters in group B provide higher scores than raters in group A.
4) We thank the anonymous reviewer for suggesting to add this alternative model results.

Acknowledgements

We would like to thank Rukiye KIZILTEPE, Duygu ESLEK, Türkan YILMAZ IRMAK, Duygu GÜNGÖR CULHA for sharing the research data with us.

Data Availability

Requests for accessing the data that support the findings of this study should be made to the corresponding author of Kızıltepe and colleagues (2022).

Declarations

Author Contribution: Ömer Emre Can ALAGÖZ – conceptualization, methodology, software, analyses, writing, editing. Yılmaz Orhun GÜRLÜK – conceptualization, software, analyses, writing, editing. Mediha KORKMAZ – conceptualization, supervising. Gizem CÖMERT – conceptualization, writing, editing, visualization.

Conflict of Interest: No potential conflict of interest was reported by the authors.

Ethical Approval: Secondary data were used in this study. Therefor ethical approval is not applicable.

References

Agresti, A. (1992). Modelling patterns of agreement and disagreement. Statistical Methods in Medical Research, 1, 201-218. https://doi.org/10.1177/096228029200100205

Ato, M., López, J. J., & Benavente, A. (2011). A simulation study of rater agreement measures with 2x2 contingency tables. Psicológica, 32(2), 385–402.

Basten, M., Tienmeier H., Althoff, R., van de Schoot, R., Jaddoe, V. W. V., Hofman, A., Hudziak, J. J., Verhulst, F. C. & Van der Ende, J. (2015). The stability of problem behavior across the preschool years: an empirical approach in general population, Journal of Abnormal Child Psychology, 44(2), 393-404. https://doi.org/10.1007/s10802-015-9993-y

Bıkmaz Bilgen, Ö. ve Doğan, N. (2017). Puanlayıcılar arası güvenirlik belirleme tekniklerinin karşılaştırılması, Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 8(1), 63-78. https://doi.org/10.21031/epod.294847

Cohen (1960). A coefficient of rater agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37-46. https://doi.org/10.1177/001316446002000104

De Los Reyes, A., Henry, D. B., Tolan, P. H. T. & Wakschlag, L. S. (2009). Linking informant discrepancies to observed variations in young children's disruptive behavior, Journal of Abnormal Psychology, 37(5), 637-652. https://doi.org/10.1007/s10802-009-9307-3

Feinstein, A. R. & Cicchetti, D. V. (1990). High agreement but low kappa: I. the problems of two paradoxes, Journal of Clinical Epidemiology, 43(6), 543-549. https://doi.org/10.1016/0895-4356(90)90158-L

Fleiss, J. L. (1971). Measuring agreement for multinomial data. Psychological Bulletin, 76(5), 378-382. https://doi.org/10.1037/h0031619

Forster, A. J., O'Rourke, K., Shojania, K. G., & van Walraven, C. (2007). Combining ratings from multiple physician reviewers helped to overcome the uncertainty associated with adverse event classification. Journal of clinical epidemiology, 60(9), 892-901.

Gisev, N., Simon Bell, J. & Chen, T. F. (2013). Interrater agreement and interrater reliability: key concepts, approaches, and applications. Research in Social and Administrative Pharmacy, 9, 330-338. https://doi.org/10.1016/j.sapharm.2012.04.004

Göktaş, A. & İşçi, Ö. (2011). A comparison of the most commonly used measures of association for doubly ordered square contingency tables via simulation. Metodoloski Zvezki, 8(1), 17-37. https://doi.org/10.51936/milh5641

Hallgren, K. (2012). Computimg inter-rater reliability for observational data: an overview and tutorial, Tutorials in Quantitative Methods for Psychology, 8(1), 23-34. https://doi.org/10.20982/tqmp.08.1.p023

Hayes, A. F. & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding, Communication Methods and Measures, 1(1), 77-89. https://doi.org/10.1080/19312450709336664

Jiang, Z. (2019). Using the iterative latent-class analysis approach to improve attribute accuracy in diagnostis classification models. Behavior Research Method, 51, 1075-1084. https://doi.org/10.3758/s13428-018-01191-0

Kızıltepe R., Eslek, D., Yılmaz Irmak, T. & Güngör, D. (2022). I am learning to protect myself with Mika:" a teacher-based child sexual abuse prevention program in Turkey. Journal of Interpersonal Violence, 37(11-12), 1-25. https://doi.org/10.1177/0886260520986272

Konstantinidis, M., Le, L. W., & Gao, X. (2022). An empirical comparative assessment of inter-rater agreement of binary outcomes and multiple raters. Symmetry, 14(2), 262. https://doi.org/10.3390/sym14020262

Kottner, J.,Audige, L., Brorson, S., Donner, A., Gajewski, B., Hrobjartsson, A., Roberts, C., Shoukri, M. & Streiner, D. L. (2011). Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. Journal of Clinical Epidemiology, 64, 96-106. https://doi.org/10.1016/j.ijnurstu.2011.01.016

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159-174. https://doi.org/10.2307/2529310

Leising, D., Ostrovski, O., & Zimmermann, J. (2013). "Are we talking about the same person here?" Interrater agreement in judgments of personality varies dramatically with how much the perceivers like the targets. Social Psychological and Personality Science, 4(4), 468-474. https://doi.org/10.1177/1948550612462414

Major, S., Seabra-Santos, M. J. & Martin, R. P. (2018). Latent profile analysis: another approach to look at parent-teacher agreement on preschoolers' behavior problems. European Early Childhood Education Research Journal, 26(5), 701-717. https://doi.org/10.1080/1350293X.2018.1522743

Miller, W. E. (2011). A latent class method for the selection of prototypes using expert ratings. Statistics in Medicine, 31(1), 80-92.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. Structural Equation Modeling, 14(4), 535–569. https://doi.org/10.1080/10705510701575396

Raykov, T., Dimitrov, D. M., von Eye, A. & Marcoulides, G. A. (2013). Interrater Agreement Evaluation: a latent variable modeling approach. Educational and Psychological Measurement, 20(10). 1-20. https://doi.org/10.1177/0013164412449016

Schuster, C. & Smith, D. A. (2002). Indexing systematic rater agreement with a latent-class model. Psychological Methods, 7(3), 384-395. https://doi.org/10.1037/1082-989X.7.3.384

Sertdemir, Y., Burgut, H. R., Alparslan, Z. N., Unal, I., & Gunasti, S. (2013). Comparing the methods of measuring multi-rater agreement on an ordinal rating scale: a simulation study with an application to real data. Journal of Applied Statistics, 40(7), 1506-1519. https://doi.org/10.1080/02664763.2013.788617

Shaffer, D., Schwab-Stone, M., Fisher, P., Cohen, P., Placentini, J., Davies, M. & Regier, D. (1993). The diagnostic interview schedule for children-revised version (DISC-R): I. Preparation, field testing, interrater reliability, and acceptability. Journal of the American Academy of Child & Adolescent Psychiatry, 32(3), 643-650. https://doi.org/10.1097/00004583-199305000-00023

Tanner, M. A. & Young, M. A. (1985). Modelling agreemet among raters. Journal of the American Statistical Association, 80(389), 175-180. https://doi.org/10.1080/01621459.1985.10477157

Thompson, D. M. (2003). Comparing SAS-based applications of latent class analysis using simulated patient classification data. The University of Oklahoma Health Sciences Center.

Uebersax, J. S. & Grove, W. M. (1990). Latent class analysis of diagnostic agreement. Statistisc in Medicine, 9(5), 559-572. https://doi.org/10.1002/sim.4780090509

Viera, A. J. & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistics, Family Medicine, 37(5), 360-363. PMID: 15883903

Von Eye, A. & Mun, E. Y. (2005). Analyzing rater agreement manifest variable methods (1st ed.). Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410611024

Yarnold, P. R. (2016). ODA vs. π and κ: paradoxes of kappa, Optimal Data Analysis, 5, 160-161. Accessed at: https://www.researchgate.net/publication/309681250_ODA_vs_p_and_k_Paradoxes_of_Kappa, 23.03.2023

Yilmaz, A. E. & Saracbasi, T. (2017). Assessing agreement between raters from the point of coefficients and log-linear models. Journal of Data Science, 15, 1-24. https://doi.org/10.6339/JDS.201701_15(1).0001

Yilmaz, A. E. & Saracbasi, T. (2019). Agreement and adjusted degree of distinguishability for square contingency tables. Hacettepe Journal of Mathematics and Statistics, 48(2), 592-604. https://doi.org/10.15672/hjms.2018.620

Zapf,A.,Castell, S., Morawietz, L., & Karch, A. (2016). Measuring inter-rater reliability for nominal data–which coefficients and confidence intervals are appropriate?. BMC medical research methodology, 16, 1-10. https://doi.org/10.1186/s12874-016-0200-9

APPENDICES

Appendix A: Latent GOLD Syntax

```
options
  algorithm
    tolerance=1e-008 emtolerance=0,01 emiterations=250 nriterations=50;
  startvalues
    seed=0 sets=10 tolerance=1e-005 iterations=50;
  bayes
    categorical=1 variances=1 latent=1 poisson=1;
  montecarlo
    seed=0 replicates=500 tolerance=1e-008;
  quadrature  nodes=10;
  missing  includeall;
  output
    parameters=first standarderrors probmeans=posterior profile bivariateresiduals;
variables
  dependent hay, ayr, cin, duy, it, miz, okul, of, yab, yet;
  latent
    Cluster nominal 2;
equations
  Cluster <- 1;
  hay <- 1 + Cluster;
  ayr <- 1 + Cluster;
  cin <- 1 + Cluster;
  duy <- 1 + Cluster;
  it <- 1 + Cluster;
  miz <- 1 + Cluster;
  okul <- 1 + Cluster;
  of <- 1 + Cluster;
  yab <- 1 + Cluster;
  yet <- 1 + Cluster;
```

Appendix B: Results for the model with three classes

We also provide the results for the model with three classes as it had slightly lower BIC value than the model with two classes. First, we investigated the class sizes. We found that the size of the first class was 0.61, the size of the second class was 0.35, and the third class was 0.04. With modal assignment, 184 subjects were assigned to class 1, 97 subjects were assigned to class 2, and only 9 subjects were assigned to class 3. These class sizes provide further reasons for sticking to the model with two classes, because the newly added class is very small; therefore, its estimates would be less accurate. Regardless of that, we provide the conditional probabilities of rating differences between raters given classes in Table B1.

Table B1:

Conditional probabilities of rating differences between different raters $Pr\left(X_{nj}^{A-B} = m \middle| Class = c\right)$

Items	Classes	$m\ (X^{parents} - X^{teachers})$									RAC
		-4	-3	-2	-1	**0**	1	2	3	4	
Item 1	Class 1	-	-	.01	.02	**.71**	.23	.03	.01	.01	**.73**
	Class 2	-	-	.01	.01	**.33**	.38	.19	.04	.04	**.28**
	Class 3	-	-	.01	.01	**.55**	.33	.09	.01	.01	**.44**
Item 2	Class 1	-	.01	-	.01	**.72**	.20	.04	.01	.01	**.73**
	Class 2	-	.01	-	.01	**.43**	.24	.12	.05	.15	**.41**
	Class 3	-	.01	-	.01	**.64**	.23	.07	.02	.03	**.44**
Item 3	Class 1	-	-	.01	.01	**.85**	.11	.02	.01	-	**.86**
	Class 2	-	-	.01	.01	**.75**	.18	.05	.01	-	**.72**
	Class 3	-	-	.01	.01	**.67**	.22	.09	.01	-	**.67**
Item 4	Class 1	.02	.06	.10	.14	**.28**	.20	.10	.04	.04	**.31**
	Class 2	.01	.03	.05	.10	**.26**	.23	.15	.09	.10	**.20**
	Class 3	.03	.09	.13	.16	**.28**	.17	.08	.03	.02	**.13**
Item 5	Class 1	.01	.01	.04	.09	**.48**	.25	.07	.03	.01	**.53**
	Class 2	.01	.01	.01	.03	**.30**	.29	.15	.12	.10	**.21**
	Class 3	.12	.15	.15	.15	**.07**	.07	.01	.01	.01	**.11**
Item 6	Class 1	-	.01	.02	.06	**.23**	.34	.18	.08	.08	**.22**
	Class 2	-	.01	.01	.01	**.07**	.20	.20	.18	.34	**.07**
	Class 3	-	.24	.18	.18	**.25**	.13	.02	.03	.01	**.44**
Item 7	Class 1	-	.01	.01	.08	**.70**	.16	.02	.01	.01	**.71**
	Class 2	-	.01	.01	.02	**.50**	.29	.11	.04	.03	**.52**
	Class 3	-	.01	.01	.06	**.68**	.20	.04	.01	.01	**.33**
Item 8	Class 1	.01	.01	.01	.04	**.62**	.22	.06	.01	.01	**.68**
	Class 2	.01	.01	.01	.01	**.33**	.28	.19	.09	.11	**.26**

	Class 3	.20	.40	.14	.12	**.13**	.01	.01	.01	.01	**.00**
Item 9	Class 1	-	.01	.01	.02	**.85**	.13	.01	.01	.01	**.88**
	Class 2	-	.01	.01	.01	**.50**	.28	.09	.07	.05	**.43**
	Class 3	-	.10	.20	.29	**.41**	.01	.01	.01	.01	**.33**
Item 10	Class 1	.01	.01	.02	.10	**.70**	.17	.01	.01	.01	**.71**
	Class 2	.01	.01	.01	.02	**.46**	.38	.05	.01	.09	**.42**
	Class 3	.01	.01	.01	.04	**.63**	.29	.02	.01	.01	**.67**