Squibs and Discussions

Identifying Sources of Disagreement: Generalizability Theory in Manual Annotation Studies

Petra Saskia Bayerl*
University of Oklahoma

Karsten Ingmar Paul†
University of Erlangen-Nuremberg

Many annotation projects have shown that the quality of manual annotations often is not as good as would be desirable for reliable data analysis. Identifying the main sources responsible for poor annotation quality must thus be a major concern. Generalizability theory is a valuable tool for this purpose, because it allows for the differentiation and detailed analysis of factors that influence annotation quality. In this article we will present basic concepts of Generalizability Theory and give an example for its application based on published data.

1. Introduction

Manual annotations are still a major source of information in many small- and large-scale projects in diverse areas of corpus and computational linguistics. Often, however, manual annotations are not reliable enough for a given application. Measures must be taken to increase and secure the consistency of linguistic annotations, if analyses and applications are not to suffer from low data quality. Because a multitude of factors may be responsible for inadequate reliability, a method is needed that is able to simultaneously consider a variety of probable factors and indicate those that are mainly responsible for low reliability in a given case. Generalizability Theory, or G-Theory (Cronbach et al. 1972), is a methodological framework specifically designed for this purpose. Because it is not restricted to any type of data or study design, it can be of great use in any kind of manual annotation project that needs to systematically identify sources of annotator disagreement. In this article we provide an outline of the approach and its basic assumptions and demonstrate its application based on an annotation study done by Shriberg and Lof (1991).[1]

* Department of Psychology, Tulsa Graduate College, 4502 East 41st Street, Tulsa, OK 74135, USA.
† Chair of Psychology, especially Organizational and Social Psychology (Prof. Dr. Moser), Lange Gasse 20, 90403 Nuremberg, Germany.

1 A more comprehensive introduction to G-Theory is provided in a longer version of this article, which is available from the authors.

The work for this article was done at the Department for Applied and Computational Linguistics, Justus-Liebig-University Giessen, Germany.

2. The G-Theory Approach

Reliability in G-Theory is defined by the amount of variation or variance observed in annotations; the lower the total variance in the data, the higher is its reliability. G-Theory further assumes that data reliability is influenced by several independent factors or *facets*, which are, individually as well as in interaction, responsible for the observed variation.[2] Sources of variation might be idiosyncratic behaviors of individual annotators or external influences like alterations in the tools used for annotations, increasing time pressure, removal or adding of rewards, or changes in the annotation scheme. Each of these influences can lead to systematic changes in an annotator's behavior and so to higher disagreement among annotators. According to G-Theory each possible facet, *annotator*, *tools*, *rewards*, and so on, will have its own independent impact on the quality, that is, reliability, of annotations. The task of a G-study is to isolate the influence of single facets and determine the degree of their impact.

2.1 Basic G-Study Designs

The main distinction with respect to G-study designs is the choice between a **crossed** and a (partially) **nested** design. In crossed designs measurements are obtained for each possible combination of facet values. Given two facets, **items** and **coders**, each individual item (phrase, phone, gesture, etc.) is annotated by all possible coders, so that each value of the *item* facet is measured on every value of the *coders* facet. Nested designs, in contrast, only measure a subset of possible combinations of facet values, for instance, when limited resources determine that only some of the coders annotate the same objects on more than one occasion. In general, fully crossed designs require a higher number of observations, but also provide more information. To obtain a full picture of possible influences crossed designs should therefore be preferred. For a detailed discussion of G-study designs, including unbalanced designs or missing data, and random and fixed facets, see, for example, Brennan (2001).

2.2 Estimating Variance Components

In fully crossed designs the total variance in the data is a result of individual facets as well as their interactions. Because G-Theory assumes independence of facets, effects of components are additive. Given three facets a, b, c, the total variance $\sigma^2(X_{abc})$ therefore is calculated as

$$\sigma^2(X_{abc}) = \sigma^2_a + \sigma^2_b + \sigma^2_c + \sigma^2_{ab} + \sigma^2_{ac} + \sigma^2_{bc} + \sigma^2_{abc,e} \quad (1)$$

where σ^2 refers to variance and the subscripts to the name of one or more facets. The subscript e in the last variance component denotes error variance. In nested designs some facets cannot be determined as independent terms due to their confounding with other facets. For instance, in a nested design with three factors a, b, c, different

2 This definition of reliability differs from the traditional true-score-model of classical reliability theory (Spearman 1904) and can be considered a modern approach to the question of consistency or 'dependability' of data measurement. A discussion of the conceptual differences is beyond the scope of this article. Information on this topic may be found in Thompson (2002) or Matt (2001).

values of *c* may be associated with different values of *b*. Here the effect for *c* will be confounded both with *bc* and the residual term *abc,e* so that no independent term for the *c* facet can be obtained. Instead of the seven variance components in the crossed design only five variance components can be calculated, again stressing the fact that nested designs provide less information than fully crossed designs

$$\sigma^2(X_{abc}) = \sigma^2_a + \sigma^2_b + \sigma^2_{ab} + \sigma^2_{c,cb} + \sigma^2_{ac,abc,e} \quad (2)$$

For information on the mathematical foundation of G-Theory and the derivation of estimates see Cronbach et al. (1972) and Brennan (2001).

2.3 Interpreting Variance Components

Based on the assumption that the total variance is a sum of single variance components, the total variance is 100%. The relative magnitude of each component with respect to the total variance is an indicator of the individual contribution of this component with respect to overall (un)reliability. A facet explaining 60% of the total variance would thus be considered a major source of variation in contrast to a minor facet explaining only 5% of the variance. For instance, given that the *coder* facet is the largest facet, variation can be explained through systematic differences in the annotation behavior of individual coders—for example, annotators differ in their tendency to set prosodic boundaries in utterances leading to systematic differences in the number of boundaries placed. In this case retraining of annotators to reach a more comparable behavior would be advisable. A high *schema* component indicates that there is systematic variation in the use of categories, whereas a high *coder–schema* interaction indicates systematic differences in annotators' use of these categories; for example, coders annotating rhetorical (RST) relations could differ in the frequency with which they use individual relations such as 'background', 'concession', 'evidence', and so forth, pointing to possible problems with the interpretation of rhetorical relations and their application. Variation mainly due to the *item* facet indicates that certain materials are harder to annotate than others. Such a result would imply retraining or elimination of overly difficult material. In consequence, the identification of distinct sources of variation should lead to specifically designed steps for improvement.

3. A Re-Analysis of Shriberg and Lof (1991)

As an illustration for the application of G-Theory we reanalyzed data provided by Shriberg and Lof (1991), who studied the accuracy of broad and narrow phonetic transcriptions. In Set A of their study they investigated four facets: *annotation scheme* (type of consonant, C), *granularity* (broad vs. narrow transcription G), *material* (continuous speech vs. articulation test, M), and *annotation team* (T). Data in Set A were given as agreement percentages. Our G-study results are shown in Table 1.

Traditionally, reliability concerns focus on disagreements among individual annotators assuming that variation is due to incommensurable annotator behavior. In our case, however, the *team* facet explains only a very small percentage of variance both as an individual factor and in interaction with other factors. This suggests that the four annotation teams are comparable in their annotation quality. The major factors responsible for the observed variance are granularity and type of consonants. Material

Table 1
G-Study results for Shriberg and Lof (1991), Table 8, Set A.

Effect	df	Variance components estimates		Percentage of total variance
Consonant (C)	23	234.86877		25.70
Granularity (G)	1	312.80278		34.23
Team (T)	3	3.70906		0.41
Material (M)	1	0.0	[−3.08672]∗	–
CG	23	99.18526		10.85
CT	69	0.0	[−8.25984]∗	–
CM	23	45.80498		5.01
GT	3	0.0	[−1.12263]∗	–
GM	1	0.0	[−6.05138]∗	–
TM	3	0.0	[−1.74740]∗	–
CGT	69	3.84207		0.42
CGM	23	111.61108		12.12
CTM	69	57.64646		6.31
GTM	3	6.04318		0.66
CGTM,e	36	38.23065		4.18
		913.74429		99.99

∗ Values set zero, original negative estimates in brackets.
For the analysis the GENOVA program as described in Brennan (2001) was used.

does not exhibit a substantial individual influence on reliability, but becomes relevant in the CGM-interaction. Our G-study therefore reveals that unreliability in Shriberg and Lof's data is caused not by idiosyncrasies of individuals, but due to the characteristics of the task, namely, granularity and scheme.

Having identified the critical facets, it might now be interesting to look at the *values* of these facets that are especially prone to produce disagreement. Because we operated with agreement data, this information can be easily obtained from the data entered into the analysis. Because neither team nor material are major sources for variance, we only have to examine the values for granularity and consonants. Due to the same reason we can base the comparison on mean values over teams and material types. For the granularity facet we find overall lower agreement in narrow transcriptions (64.15%) compared to broad transcriptions (89.46%). On the consonant facet we can differentiate critical phonemes such as /ð/ or /ʃ/ from uncritical ones (e. g., /j/, /b/). Interpreting the CG-interaction in this light, disagreement on consonants in narrow transcriptions seems to be comparably higher than in broad transcriptions. Implications from this study would be that the selection of annotators and the training of annotation teams are successful in producing comparable results. For high reliability, however, transcriptions should be done on a broad level with specific training for difficult consonants and some special care for material from articulation tests (see CGM-interaction).

4. Practical Considerations in Planning G-Studies

In planning and conducting a G-study some deliberation is necessary to achieve interpretable results. Foremost, the overall quality of the G-study depends on the choice of factors that completely and accurately represent the situation of the annotators. As it is

quite easy to overlook relevant but rather inconspicuous factors like minor changes in the annotation tool or increasing time pressure due to upcoming project deadlines, the choice of correct facets relies heavily on the experience and knowledge of the researcher. The statistical results, however, will give indications for likely misspecifications of facets by showing a high error or rest variance σ_e for the tested model. Theoretically, the number of facets that can be included in a G-study is unlimited. Having more than four or five facets in one study might make the final interpretation overly complex, however. Even though there is no minimum necessary number of observations, missing data due to a low number of observations pose a problem for model interpretation. Approaches to deal with such unbalanced designs are given by Brennan (2001) and Chiu and Wolfe (2003). Additionally, there is no clear-cut rule when a component might be considered 'too small' to be of importance. As a rule of thumb a component of less than 8% might be considered 'small', but the decision remains one of 'relative importance' depending on the distribution of explained variance across components.

5. G-Theory and Agreement Indices

Two well-known measures for capturing the quality of manual annotations are agreement percentages and the kappa statistic (Cohen 1960; Carletta 1996; Eugenio and Glass 2004). Both measures provide a "summary index" (Agresti 1992) that expresses the degree of (dis)agreement among coders. Where the calculation of percentages and kappa provides a measure for overall reliability (or reliability indices for individual facets), G-Theory has been designed to analyze multiple possible influencing factors in a single run and to compare the relative importance of components among each other. Shriberg and Lof (1991), for instance, compare narrow and broad transcriptions using graphics that show the agreement percentages for each consonant for both transcription types, arguing that overall narrow transcriptions seem unreliable. Based on their experience with the study context they further assumed that these differences were not due to annotator training, behaviors, or experience. They did not provide any direct evidence, however. The G-study presented in this article could prove both assumptions in a single run. It further allows us to investigate the interactions and mutual influences of these factors, thus clearly exceeding the possibilities of summary statistics. As we have seen in the example, G-Theory, however, does not provide answers as to which values of a facet are responsible for higher or lower reliability. This information must be obtained by a review of the data. Agreement indices and G-Theory should thus not be seen as competing, but rather as complementary, approaches. Kappa can serve as a first approximation to the degree of disagreement present in the data, whereas G-Theory in a second step investigates the underlying reasons of inadequate reliability and subsequently guides efforts to improve reliability.

6. Final Remarks

Generalizability theory is a valuable approach for identifying problematic areas in annotation projects. The investigation of multiple facets at the same time can provide a clearer understanding of reasons underlying insufficient annotation quality and subsequently offer avenues to its improvement. In this article we could not give more than a passing glance over the possibilities provided by the G-Theory approach. For the interested reader, Shavelson and Webb (1981) give a good introduction into the material. Further references are provided throughout the article and in the reference section.

References

Agresti, Alan. 1992. Modelling patterns of agreement and disagreement. *Statistical Methods in Medical Research*, 1(2):201–218.

Brennan, Robert L. 2001. *Generalizability Theory*. Statistics for Social Science and Public Policy. Springer Verlag, New York.

Carletta, Jean. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Chiu, Christopher W. T. and Edward W. Wolfe. 2003. A method for analyzing sparse data matrices in the generalizability theory framework. *Applied Psychological Measurement*, 26(3):321–338.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Cronbach, Lee J., Goldine C. Gleser, Harinder Nanda, and Nageswari Rajaratnam. 1972. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. John Wiley & Sons, Inc., New York.

Eugenio, Barbara Di and Michael Glass. 2004. The kappa statistic: A second look. *Computational Linguistics*, 30(1):95–101.

Matt, Georg E. 2001. Generalizabilty theory. In Neil J. Smelser and Paul B. Baltes, editors, *International Encyclopedia of the Social and Behavioral Sciences*. Elsevier, Oxford.

Shavelson, Richard J. and Noreen M. Webb. 1981. *Generalizability Theory: A Primer*. Sage Publications, Newbury Park, London, New Delhi.

Shriberg, Lawrence D. and Gregory L. Lof. 1991. Reliability studies in broad and narrow phonetic transcription. *Clinical Linguistics and Phonetics*, 5(3):225–279.

Spearman, Charles. 1904. The proof and measurement of association between two things. *American Journal of Psychology*, 15:72–101.

Thompson, Bruce. 2002. A brief introduction to Generalizability Theory. In Bruce Thompson, editor, *Score Reliability: Contemporary Thinking on Reliability Issues*. Sage Publication, Thousand Oaks, CA, pages 43–58.

Statistical Software

- GENOVA, urGENOVA, mGENOVA: Available online at http://www.education.uiowa.edu/casma/GenovaPrograms.htm