Inductive Event Data Scaling using Item Response Theory [*]

Philip A. Schrodt
Dept. of Political Science
University of Kansas
Lawrence, KS 66045
785.864.9024 (phone) - 785.864.5700 (fax)
Email: schrodt@ku.edu

Version 1.0B1 : July 17, 2007

[*] Paper prepared for delivery at the Summer Meeting of the Society for Political Methodology, Pennsylvanian State University, 18 - 20 July 2007. An earlier version was presented at the 2007 Annual Meeting of the International Studies Association, Chicago. This research was supported in part by the National Science Foundation Human and Social Dynamics Initiative SES-0527564. The data sets discussed in this paper, as well as a pdf version of the paper with color graphics, can be downloaded from the KEDS project web site: `http://web.ku.edu/keds/`.

Abstract

Political event data—categorical data showing who did what to whom (and when) derived from news reports—are frequently converted to an interval level measurement by assigning a numerical scaled value to each event. All of the existing scaling systems rely on non-replicable expert assessments to determine these numerical scores, which do not take into account the characteristics of the data that will be aggregated. This paper uses item response theory (IRT)—a technique originally developed for the scaling of test scores—to derive scales inductively, using event data on Israeli interactions with Lebanon and the Palestinians for 1991-2007. In the IRT model, the probability of an event being reported in an interval of time by a specific news source is modeled as a logistic function on a latent dimension determined from the data itself. Monthly scores on this latent trait are calculated using three IRT models: the single-parameter Rasch model, and two-parameter models that add discrimination and guessing parameters. The three formulations produce generally comparable scores (correlations around 0.90 or higher). The Rasch scales are less successful than the expert-derived Goldstein scale in reconciling the somewhat divergent sets of events derived from the *Agence France Presse* and Reuters news services. This is in all likelihood due largely to the low weighting given uses of force by the IRT models, because force events are common in these two data sets. A factor analysis of the event counts shows that a single cooperation-conflict dimension generally accounts for about two-thirds of the variance in these dyads, but a second case-specific dimension explains another 20%. Finally, moving averages of the scores generally correlate well with the Goldstein values, suggesting that IRT may provide a route towards deriving a purely inductive (and hence replicable) scale.

1 Introduction

Political event data—categorical data showing who did what to whom (and when) derived from news reports—are frequently converted to an interval level measurement by assigning a numerical scaled value to each event. In international relations, popular unidimensional conflict-cooperation scales include those of Azar (Azar and Sloan 1975; Azar 1980, 1982) for the Conflict and Peace Data Bank (COPDAB) event coding system, and Vincent (1979) and Goldstein (1992) for World Events Interaction Survey (WEIS) events; more recently, VRA Inc. (`http://vranet.com/`) has developed a three-dimensional scale for the Integrated Data for Events Analysis (IDEA) coding system. All of these coding systems rely on expert assessments to assign the numerical scores, and those assessments do not take into account the actual data that will be aggregated, instead relying on some abstract generalization (or wild-assed guess) of the appropriate value. Consequently, for example, in COPDAB "three riots equals one thermonuclear war."

A second problem that has emerged in event data coding is the lack of correspondence between multiple sources of events, for example the Reuters and *Agence France Presse* (AFP) wire services, and *The New York Times* and *Washington Post* newspapers. All news services report only a fraction of the total number of occurrences of possible codeable events, and differ in how they do this. For example it appears that Reuters is more likely than AFP to report on economic events; AFP is more likely to report an on-going story in small fragments. News agencies differ in the number of personnel they assign to different geographical areas, which in turn affects the number of stories they will file. The likelihood of reporting may also differ depending on the event, for example with biases towards events that are either very "newsworthy" (suicide bombings) or easy to report (news conferences).

This paper uses item response theory (IRT)—a technique originally developed for the scaling of test scores—in an effort to resolve both issues, using event data from Israel's interactions with the Palestinians and with Lebanon for 1991-2007 coded with WEIS and the Conflict and Mediation Event Observations (CAMEO; Gerner et al 2007) coding systems. In the IRT model, the probability of an event being reported in an interval of time for a particular dyad (for example Israel and Lebanon) and news source is modeled as a logistic function on a unobserved latent trait. The score on this trait can then be computed as a function of these probabilities, with rare events indicating that a higher score on the latent trait than common events. These scores will be compared to those obtained with the Goldstein scale, as well as determining whether IRT-derived scales can be used to normalize events derived from different news sources. In addition, a factor analysis will be done to determine the extent to which the unidimensionality of the latent trait is appropriate.

2 Scaling Event Data

All event data begin as categorical data: reports on interactions between two political actors are evaluated by a human coder or an automated system, and assigned to one of a number of possible categories, ranging from about 15 categories in the CODAB system to more than 100 categories in the Behavioral Correlates of War (BCOW; Leng 1987) and some of the earlier versions of IDEA. The original WEIS system contained about 66 categories; contemporary systems (IDEA and CAMEO) contain about 100, organized into a three-level hierarchy with about 20 categories at the highest level.

Unfortunately, the methods available for analyzing sequences of categorical data are relatively limited, and most occur in fields quite distant from political science, notably biology (DNA and amino acid sequences) and linguistics (phonology, the study of the sequences of sounds). On the other hand, an huge literature exists for the study of interval-level time series data, and much of those methods have been developed in economics, a close cognate to political science. An intermediate approach exists—using the count of the events—but the techniques available in this approach are still relatively limited compared to those of interval-level time series, and event count models have become common in political science only in the past couple of decades.

The tendency to want to apply interval-level methods on this nominal-level data meant that scales have been used to aggregate event data since almost the beginning of their development. Scaling has also been *controversial* from almost almost the beginning: Goldstein (1992: 372-374) provides an extended and rather illuminating survey of this debate as it existed fifteen years ago. Of the two most widely used system, COPDAB and WEIS, the COPDAB system was designed from the beginning to be scaled, and a scale is incorporated into COPDAB itself (Azar and Sloan 1975; Azar 1982), while McClelland (1983) opposed the use of scales for WEIS. McClelland's sentiments notwithstanding, a number of scales for WEIS were developed—Goldstein (1992) discusses five such efforts. Prior to Goldstein's efforts, Vincent's (1979) scale seems to have been the most widely used.

Enter Goldstein (1992). In a small but remarkably influential exercise, Joshua Goldstein produced a simple scale for WEIS events, anchored at -10 for the conflictual WEIS category 223 ("military attack"), moving to a neutral 0.0 for category 025 ("explain or state policy; state future position") and going to 8.3 for the cooperative category 072 ("extend military assistance"). These "Goldstein scores," as they have become known in the field, are now used almost exclusively for the scaling of WEIS data. The *Web of Science* database—the database formerly known as the *Social Science Citation Index*—lists 68 citations to this article from its publication to the present: most of these presumably are from articles using the scale in statistical studies. A further substantial unpublished literature using the Goldstein scale exists in conference papers.

Yet despite the ubiquity of Goldstein scores in the literature, the expert panel that produced them was quite *ad hoc.* Goldstein, at the time an assistant professor at the University of Southern California, asked 20 of his colleagues to assign numerical scores, with -10 defined as the value of the most conflictual event, 0 as neutral, and 10 as the value of most cooperative event. Only 8 individuals agreed to do this—unsurprisingly, they were disproportionately Goldstein's fellow assistant professors—and the published scores are simply the mean values of those eight responses. These weights are generally comparable to those of Vincent, though Goldstein finds the new scale " 'performs' slightly better than the Vincent scale in distinguishing the fine details of structure" (pg. 382) in three studies Goldstein replicated using both scales. Goldstein viewed this exercise as largely a corroboration of Vincent's scale rather than a replacement for it; in all likelihood the reason that the Goldstein scale became so widely used is that it was published in the readly-read and readily-accessible *Journal of Conflict Resolution* while Vincent's efforts were only available in a relatively obscure monograph.

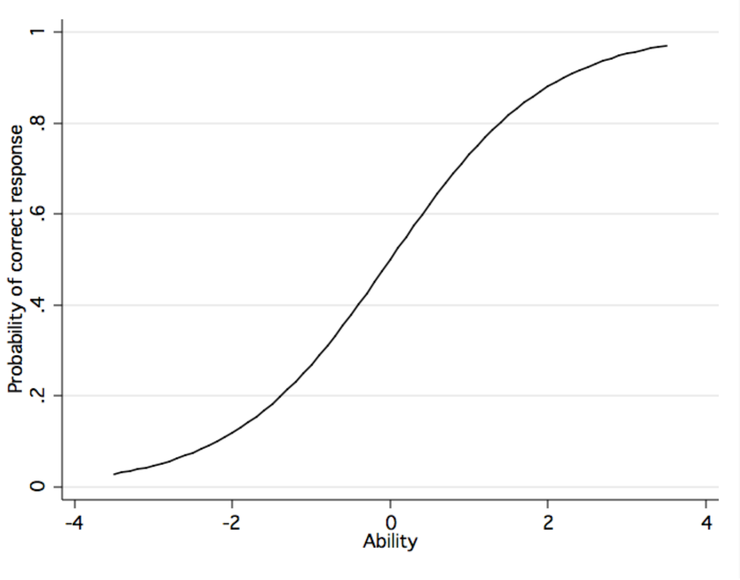

Figure 1: Example of logistic ability-response curve

In the early years of the 21st century, a second much more elaborate effort was made to develop a scale for the IDEA system (Bond et al 2003). This involved a much larger panel of experts and a web-based interface that allowed this panel to be more geographically and culturally diverse. Unfortunately, the project did not appear to reach completion.[1] Finally, Shellman (2004) has developed a scale for internal events coded by the Intranational Political Interactions (IPI) project using techniques similar to those of Goldstein, but with a larger 17-person expert panel, and a more sophisticated method of aggregating the panel's scoring.

3 Item Response Theory

Item response theory (IRT; Embretson and Reise 2000, Baker and Kim 2004) was originally developed for standardizing the scoring of tests. The underlying model is that questions on a test differ with respect to both their difficulty and their ability to differentiate among students of varying abilities. At the core of IRT is the assumption that each question can be associated with a logistic equation

$$\frac{1}{1+e^{-\alpha_i(\theta-\beta_i)}} \tag{1}$$

that associates the latent (unobserved) "ability" variable θ with the probability of a correct response to the question, as shown in Figure 1.

In the equation, β_i is proportional to "difficulty", specifically the ability level corresponding to a 50% likelihood of getting the answer correct. α_i is proportional to the degree of discrimination for a specific question i—the larger the value of α, the steeper the transition between the probability of a correct and incorrect answer, and hence the greater the level of discrimination.

[1] If you have information to the contrary, please let me know...

Birnbaum (1968) extended this model by incorporating a "guessing" parameter c_i to incorporate the possibility that low-ability students might select the correct response by chance, leading to the three-parameter model

$$c_i + (1 - c_i)\frac{1}{1 + e^{-\alpha_i(\theta - \beta_i)}} \tag{2}$$

Because "ability" is a latent variable, the challenge of IRT estimation is to generate a set of α_i, β_i and c_i that are consistent with an underlying but unobserved variable. A variety of methods exist for doing this, and conveniently these have now been implemented in various statistical packages, including Stata, SAS and R. For this paper, I have used two statistical packages. First, the parameters of the Rasch model were estimated using `raschtest` Stata module developed by Jean-Benoit Hardouin for the "Free IRT Project" and posted at `http://anaqol.org/index.php`. The Rasch and multi-parameter models were estimated using Dimitris Rizopoulos's recent R package `ltm` (Rizopoulos 2006) which estimates a variety of latent trait models based on the IRT approach. Estimates for the full 3-parameter Birnbaum model would not converge for this data, but I was able to estimate two-parameter models involving difficulty and discrimination (α_i and β_i) using the `ltm` command `ltm(data.frame ∼ z1)` and a model involving difficulty and guessing (c_i and β_i) using the `ltm` command `tpm(data.frame, type = "rasch", max.guessing = 1)`.

Both packages provide a variety of different options for estimation and optimization; the default values were used in this analysis. The estimation involves a numerical solution of a conditional maximum likelihood function and is somewhat computationally intensive, though well within the capabilities of a contemporary personal computer: estimation of the 18-item, 180 case model analyzed here required about a minute on a rather antiquated 1.6 GHz Apple Macintosh G5 using a PowerPC processor.

While the domains of educational testing and international event data might first seem quite disparate, the IRT approach in fact solves exactly the event data scaling problems discussed above. Specifically, all of the event scales except for the multi-dimensional VRA system assume that behavior can be summarized on a unidimensional scale, generally assumed to be some function of a conflict-cooperation continuum. The probability of a particular event being generated depends on where the interaction is on that scale (or, in practice, observed events are used to determine the position of the dyad on the scale), which, following the IRT approach, it is certainly reasonable to assume would follow a logistic curve. Furthermore we can assume that different news sources would have different values of α_i, β_i and c_i associated with them, depending on the likelihood of generating events of a particular type. The IRT analysis, therefore, will provide both an estimate of the response curve of each event associated with a source, and, based on the occurrence of set of events, an overall estimate of the magnitude of the dyadic interaction on the underlying latent dimension. These scores, in turn, can be used to generate a time series of the intensity of interaction over time.

This approach provides at least three advantages over the existing method of expert-derived scales. First, it is systematic and replicable, whereas the existing techniques are dependent on the non-reproducible judgments of a set of experts. Second, it can adjust for the observed characteristics of a specific news source, which is not the case with any of the expert-derived measures. Third, it can be done quickly and inexpensively, which is

Table 1: Mean frequency of 1s in monthly data by cue category

CAMEO Category	AFP PSE	REU PSE	AFP LBN	REU LBN
01	0.734	0.622	0.310	0.228
02	0.462	0.253	0.092	0.073
03	0.554	0.409	0.179	0.083
04	0.810	0.715	0.315	0.249
05	0.457	0.352	0.109	0.073
06	0.000	0.000	0.005	0.000
07	0.071	0.062	0.060	0.047
08	0.620	0.446	0.277	0.202
09	0.092	0.062	0.022	0.021
10	0.261	0.181	0.054	0.047
11	0.413	0.280	0.120	0.140
12	0.266	0.207	0.114	0.062
13	0.250	0.150	0.109	0.057
14	0.141	0.057	0.033	0.021
15	0.054	0.010	0.065	0.031
16	0.207	0.212	0.043	0.047
17	0.603	0.513	0.255	0.197
18	0.386	0.109	0.201	0.150
19	0.728	0.585	0.620	0.575
20	0.000	0.000	0.000	0.000

particularly important if one is doing dynamic development of new coding schemes such as IDEA and CAMEO.

4 Data

The data used in this study were coded into both the WEIS (McClelland 1976) scheme and the CAMEO scheme (Gerner et al 2002, 2005, 2007) using TABARI, a computer program that creates event data from machine-readable text.[2] The events were coded from two sources:

- Reuters News Service lead sentences obtained from the NEXIS data service for the period May 1991 through May 1997, the Reuters Business Briefing service for June 1997 through September 1999, and the Factiva data service for October 1999 through January 2007.

[2] Discussions of machine coding can be found in Gerner et al. (1994), Schrodt and Gerner (1994), and Bond et al. (1997) and King and Lowe 2003. The codebook for CAMEO and the TABARI program can be downloaded from `http://web.ku.edu/keds/data.dir/cameo.html`

- *Agence France Presse* (AFP) lead sentences obtained from the NEXIS data service for May 1991 through January 2007, with the exception of about ten months, randomly dispersed throughout the first six years of the data, when no AFP stories are available.[3]

The coding software, coding dictionaries and data are available at the KEDS web site, `http://web.ku.edu/keds`. Following the standard practice of the KEDS project, the event data were run through a "one-a-day filter" that discarded multiple instances of any source-target-event occurring in a single day: this eliminates duplicate and developing stories, albeit at the cost of eliminating some true multiple events, particularly acts of violence during high conflict periods.

The analysis considers two dyads: Israeli actions towards the Palestinians, and Israeli actions towards Lebanon. For the Palestinians this includes actions towards both state and sub-state actors having PSE or PAL as the primary code in the CAMEO actor coding system[4]; for Lebanon it includes any actor having LBN as the primary code: details on the CAMEO actor coding system can be found in Gerner et al (2005, 2007). These are both relatively dense dyads: the Israel→Palestinian dyad contains 13,138 events in the AFP series and 7,623 events in the Reuters series; the Israel→Lebanon dyad contains 2,236 events in the AFP series and 1,638 events in the Reuters series. Since Israel is the source in both dyads, I will simply refer to the dyads by the identity of the target.

The data were aggregated by month using first the Goldstein scale and then the scales determined by the IRT estimation. The Goldstein scores are the total of the scores of all of the events by month; these were produced using the `KEDS_Count` utility program available on the KEDS project web site. Because of the missing stories in AFP, the results of the two series were aligned by deleting months in Reuters where there was no corresponding data in AFP, consequently there are only 180 observations in the final data set. The time series plots in Figures 4 through 11 show these as continuous series rather than including the gaps (which would be invisible at this scale anyway).

The data used to derive the IRT scales required more extensive transformations. IRT scaling assumes a dichotomous response—in the usual application of testing, the response is 1 if an answer is correct and 0 if it is incorrect.[5] In order to convert the event data stream to this format, a dichotomous score was produced for each two-digit CAMEO category (see Appendix) for each month depending on whether the number of events was above or below the mean monthly frequency of that event type. Because of their very low frequencies in these dyads, categories 06 ("material cooperation") and 20 ("unconventional mass violence") were not included in the IRT estimation.

This method of reducing the data is not the only possible way to produce a dichotomous measure, and a couple of points about it are worth noting. First, it results in a *substantial*

[3] The months removed from the data set due to missing data are Feb-92, Mar-92, Aug-92, Sep-92, Nov-92, Oct-95, Jul-98 and Aug-98.

[4] PSE is the ISO-3166-1 Alpha3 code for "Palestinian Occupied Territories." In the CAMEO actor coding framework, we use this for Palestinian quasi-state entities, the Palestine Liberation Organization and the post-Oslo Palestinian Authority. PAL refers to ethnic Palestinians and includes the actions of Palestinian non-state actors such as Hamas and Islamic Jihad. LBN in the ISO-3166-1 code for Lebanon, and replaces the LEB used in our earlier work (and WEIS). ISR is the same in both systems. For further information see `https://cameocodes.wikispaces.com/`

[5] Polytomous IRT methods exist and I may pursue these in the future.

reduction in the information provided by the data stream. Second, the frequency of monthly event data is generally skewed, usually with the mean will be greater than the median, and consequently in most instances there will be more 0s than 1s, as Table 1 shows. Third, and perhaps most significantly, because every category is recoded with respect to its own mean, this measure eliminates the differences in the absolute frequencies of the various categories.

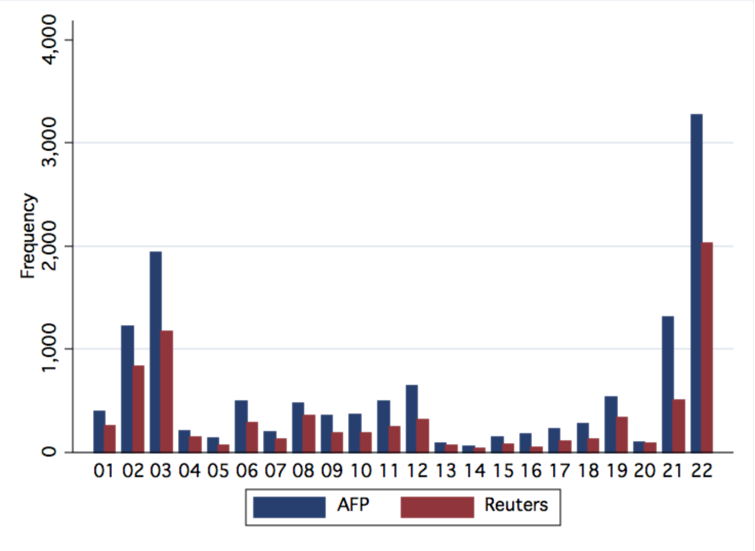

Figure 2: Frequency of WEIS cue category events in AFP and Reuters having Israel as actor and Palestinians as target

5 Comparison of AFP and Reuters

The first test of the IRT scaling will be to determine whether it can be used to provide an adjustment for the different category frequencies found in AFP and Reuters. Figures 2 through 7 show various measures of the differences between these two data sets using the WEIS coding scheme. Figures 2 and 3 show the frequencies of the WEIS cue categories for AFP and Reuters for the Palestinian and Lebanon cases; Figures 4 and 5 show the total number of events by month; and Figures 6 and 7 show the total Goldstein scores by month.

While the various measures are certainly well-correlated, they are by no means identical, nor one would expect them to be. The two news services have different editorial policies, correspondents in different locations, and different competing stories. Furthermore, because Israel, Palestine and Lebanon are geographically compact, well-monitored areas where journalists can usually move around with relatively little danger (compared, say, to contemporary Iraq or Liberia in the early 1990s) these comparisons are among the closest that one is likely to get. Even here, for example, one finds AFP reporting about 50% more events in the Israel→Palestinian dyad than Reuters reports, though this ratio is not constant: for example in the WEIS category 21 ("seize"), AFP has more than twice as many events. However, these differences are substantially less in the Israel→Lebanon dyad, and in one category, WEIS 12 ("accuse"), Reuters actually generates slightly more events than AFP.

Figures 8 and 9 show the results of the Rasch scaling for the Palestinians and Lebanon respectively.[6] The default Rasch estimation in `raschtest` assigns a "difficulty" value of 0.0 to the category 19 and the other values are estimated relative to this, almost all of the difficulty parameters were positive, since 19 is one of the most frequent, and hence "easiest" categories. Consequently the scaled scores are positive, rather than negative, although they do appear to measuring a cooperation-conflict dimension, an issue I will discuss in more

[6] These figures are produced using the `genscore` variable in the Hardouin `raschtest` Stata routine.

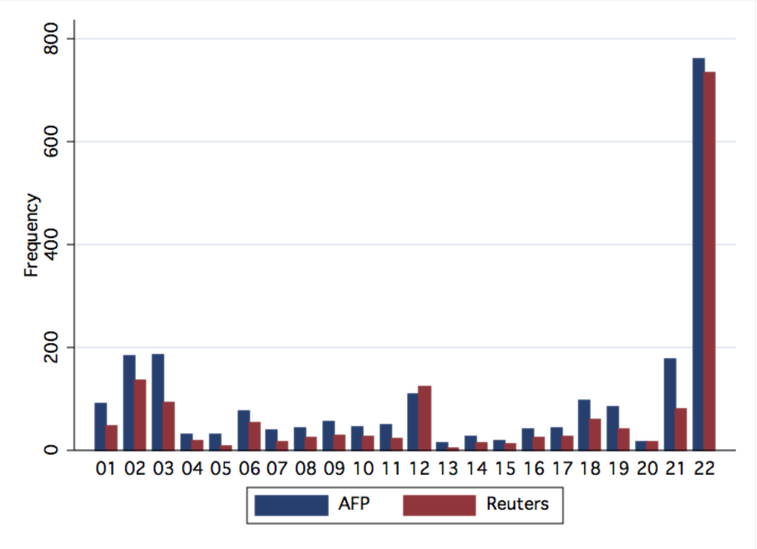

Figure 3: Frequency of WEIS cue category events in AFP and Reuters having Israel as actor and Lebanon as target

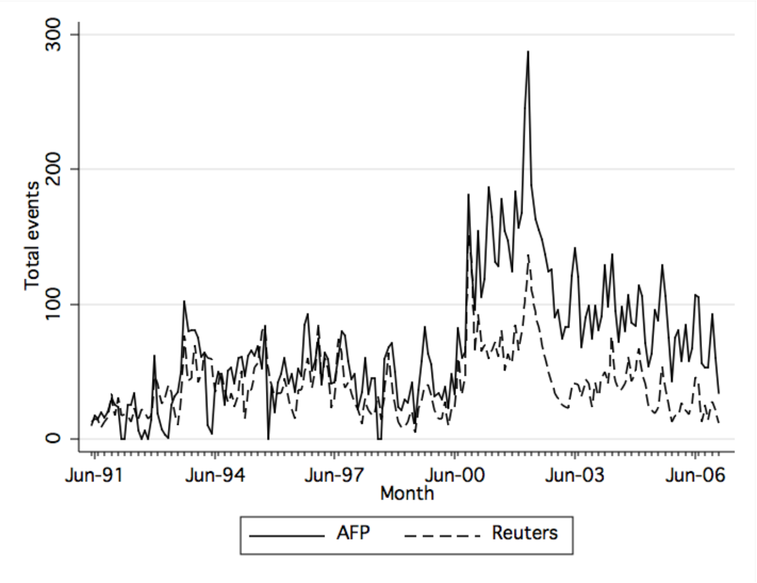

Figure 4: Total events by month for AFP and Reuters having Israel as actor and Palestinians as target

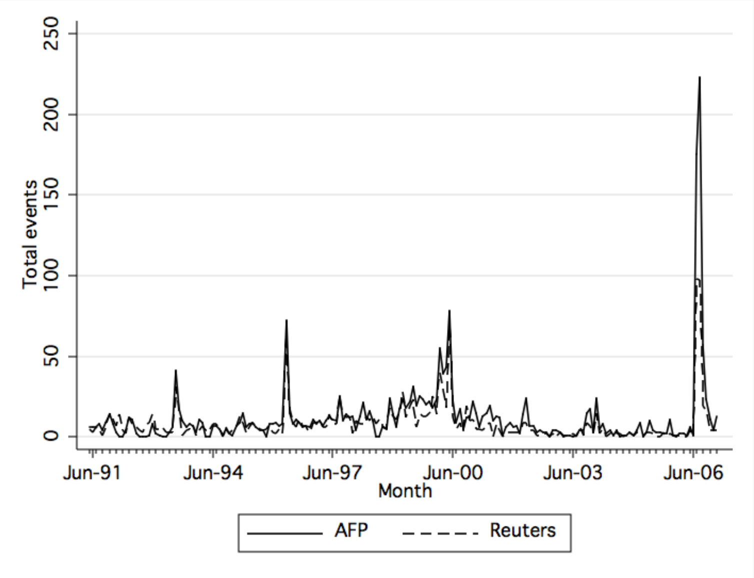

Figure 5: Total events by month for AFP and Reuters having Israel as actor and Lebanon as target

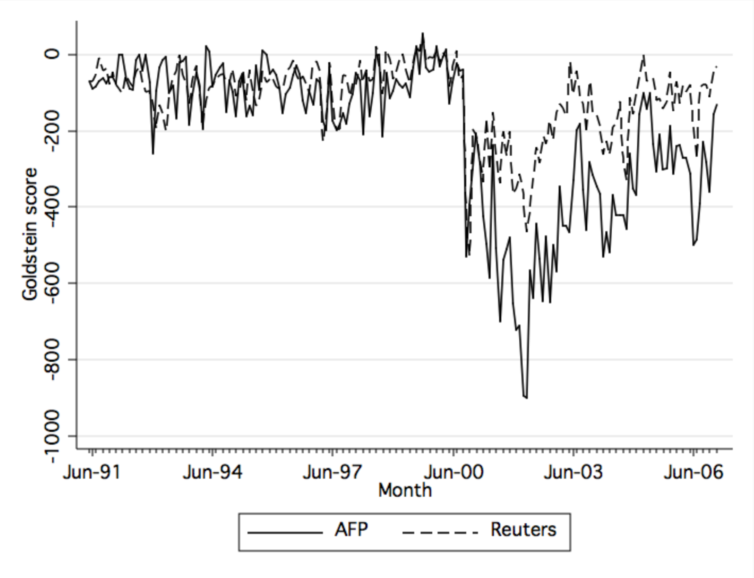

Figure 6: Monthly Goldstein scale scores for AFP and Reuters having Israel as actor and Palestinians as target

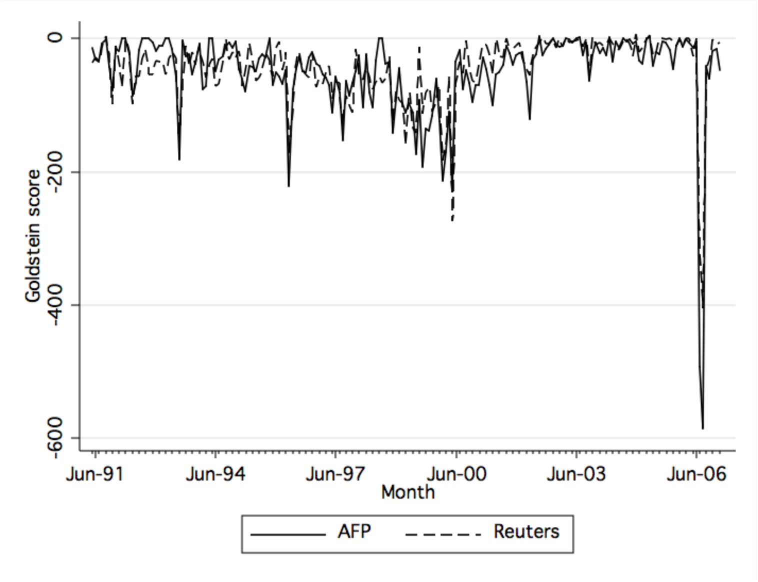

Figure 7: Monthly Goldstein scale scores for AFP and Reuters having Israel as actor and Lebanon as target

detail in the section 7.

While the AFP and Reuters curves in figures 8 and 9 are very closely aligned during many periods, they also depart substantially in others. In the Palestinian case, the AFP intensity measure is about 50% higher than the measure for Reuters during the first two years of the *al-Aqsa intifada* that began in September 2000, and AFP records a negative intensity score in the first part of the series during the period leading up to the Oslo agreement, whereas the Reuters measure rises slightly during this period. The Lebanon also shows a number of periods when the two series diverge by about 0.5, particularly in the first half of the series, prior to the Israeli withdrawal from Lebanon.

These eye-ball assessments are confirmed in the correlation matrices reported in Tables 2 and 3, which considers both the Rasch estimator and the two two-parameter alternatives. While the correlations of the AFP and Reuters Rasch series are certainly significant—0.49 for Palestine and 0.62 for Lebanon—those correlations are substantially lower than the correlations for the Goldstein scaled scores—0.80 for Palestine and 0.89 for Lebanon. Consequently the original objective of using IRT to find a superior method of reconciling multiple sources does not appear to have been achieved.[7]

Two-parameter models involving the "discrimination" parameter α_i and the "guessing" parameter c_i were estimated using the Rizopoulos R routine.[8] The inclusion of the c_i param-

[7] It is possible, but unlikely, that some of this difference is due to the Rasch scores being computed on CAMEO data and the Goldstein scores on WEIS. However, Gerner et al 2002 show a close correspondence between the two systems, and many of the distinctions are either in low-frequency categories such as "warn" and in the 3-digit and 4-digit subdivisions, so it seems unlikely that these distinctions would account for the 30% difference in variance explained.

[8] Rizopoulos's `factor.scores()` procedure produces a score for each unique *pattern* of scores, but unlike `raschtest`, does not produce these for each case. This disparity was resolved by the use of a little `perl` program that read each case in the original data, found the equivalent pattern in the `factor.scores()`

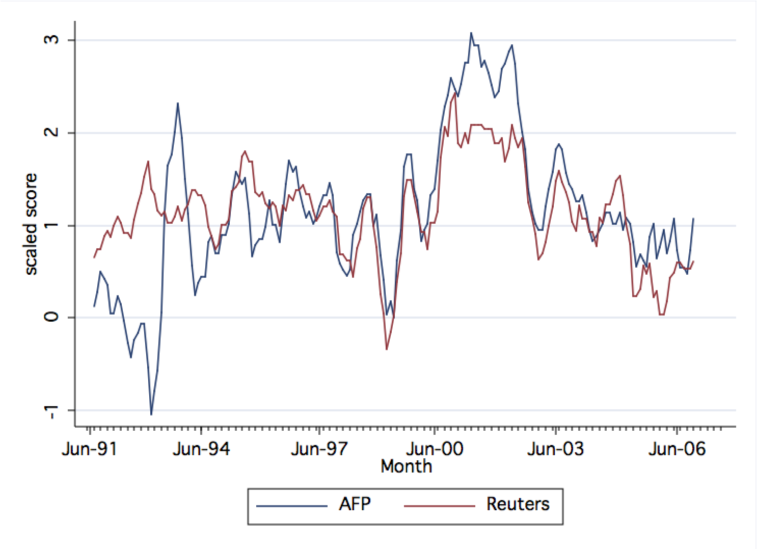

Figure 8: Monthly Rasch scores for AFP and Reuters having Israel as actor and Palestinians as target, 5-month moving average

eter is of particular interest, since this would be equivalent to the likelihood that an event would be reported when there is no conflict in the dyad (assuming that the latent dimension is in fact cooperation-conflict) or, in a sense, the level of random "background" of various event types for the two news services.

Despite the promise of the two-parameter models, the correlations in Tables 2 and 3 show that the additional parameters in fact make almost no difference. Most of resulting series correlate at levels in excess of 0.95; the lowest correlations (still in the range 0.85 - 0.97) are between the Rasch model and the two-parameter guessing model. Tests using the `anova()` procedure indicate that most of the pairs of series are significantly different at at least the 0.05 level, but the correlations between the series are so high that this clearly is not going to make any major difference in the pattern of the resulting scores. Finally, the correlations between the AFP and Reuters series for the 2-parameter models are only about 0.05 higher for the Palestine series and not higher at all for the Lebanon series, so these do no better at reconciling the two series.

output, and then wrote the score for that pattern to a file. Note—reassuringly—that the Hardouin and Rizopoulos results for the Rasch model correlate at 0.99; a scatterplot shows the slight differences occur at the extreme tails, and are presumably accounted for by differences in the numerical optimization.

Table 2: Correlations of count, Goldstein and factor score measures for AFP and Reuters, Israel actions to Palestinians (N=180)

	ispsafn	ispsren	ispsafgs	ispsregs	scafppse	screupse	psafrasc	psafdisc	psafgues	psrerasc	psredisc
ispsren	0.7459										
ispsafgs	-0.8477	-0.5154									
ispsregs	-0.7252	-0.7202	0.8053								
scafppse	0.7000	0.5652	-0.4612	-0.3770							
screupse	0.4814	0.6965	-0.2633	-0.4293	0.4944						
psafrasc	0.7138	0.5927	-0.4675	-0.4081	0.9898	0.5295					
psafdisc	0.6814	0.5767	-0.4353	-0.3754	0.9594	0.5272	0.9750				
psafgues	0.6914	0.5845	-0.4528	-0.3986	0.9445	0.5277	0.9666	0.9923			
psrerasc	0.5015	0.7081	-0.2744	-0.4410	0.5041	0.9920	0.5407	0.5374	0.5397		
psredisc	0.4351	0.6727	-0.1453	-0.3084	0.4985	0.9125	0.5358	0.5464	0.5464	0.9203	
psregues	0.4256	0.6582	-0.1250	-0.2865	0.4965	0.8805	0.5351	0.5455	0.5449	0.8944	0.9881

Variables:

Event counts: ispsafn (AFP); ispsren (Reuters)

Goldstein scores: ispsafgs (AFP) ; ispsregs (Reuters)

Rasch factor scores, Hardouin software: scafppsn (AFP), screupsn (Reuters)

Rasch factor scores, Rizopoulos software: psafrasc (AFP), psrerasc (Reuters)

Factor scores for difficulty and discrimination model : psafdisc (AFP), psredisc (Reuters)

Factor scores for difficulty and guessing model : psafgues (AFP), psregues (Reuters)

Table 3: Correlations of count, Goldstein and factor score measures for AFP and Reuters, Israel actions to Lebanon (N=180)

	islbafn	islbren	islbafgs	islbregs	scafplbn	screulbn	lbafrasc	lbafdisc	lbafgues	lbrerasc	lbredisc
islbren	0.9227										
islbafgs	-0.9209	-0.8986									
islbregs	-0.8311	-0.9222	0.8974								
scafplbn	0.6766	0.6612	-0.6769	-0.6353							
screulbn	0.5589	0.6509	-0.5754	-0.6214	0.6338						
lbafrasc	0.6884	0.6830	-0.6813	-0.6487	0.9904	0.6455					
lbafdisc	0.6714	0.6657	-0.6575	-0.6305	0.9799	0.6271	0.9910				
lbafgues	0.6722	0.6707	-0.6550	-0.6384	0.9481	0.6193	0.9707	0.9856			
lbrerasc	0.6341	0.7224	-0.6337	-0.6683	0.6637	0.9824	0.6840	0.6660	0.6610		
lbredisc	0.5878	0.6839	-0.5914	-0.6263	0.6270	0.9422	0.6518	0.6348	0.6344	0.9622	
lbregues	0.4269	0.5219	-0.4592	-0.4991	0.5261	0.8551	0.5441	0.5266	0.5249	0.8565	0.9142

Variables:

Event counts: islbafn (AFP); islbren (Reuters)

Goldstein scores: islbafgs (AFP) ; islbregs (Reuters)

Rasch factor scores, Hardouin software: scafplbn (AFP), screulbn (Reuters)

Rasch factor scores, Rizopoulos software: lbafrasc (AFP), lbrerasc (Reuters)

Factor scores for difficulty and discrimination model : lbafdisc (AFP), lbredisc (Reuters)

Factor scores for difficulty and guessing model : lbafgues (AFP), lbregues (Reuters)

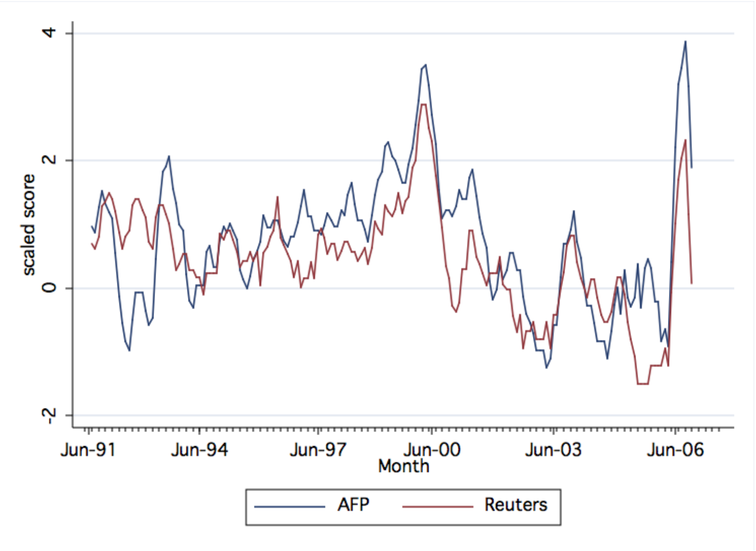

Figure 9: Monthly Rasch scores for AFP and Reuters having Israel as actor and Lebanon as target, 5-month moving average

6 Dimensionality

A second question relevant to the IRT approach is whether the event data actually do have a single latent dimension. In order to explore this, I did a simple factor analysis with the Stata `factor` routine using the principal factor estimation method on the monthly count of events in each of the CAMEO categories; categories 06 and 20 were against excluded because there were no events in the category. The results of this analysis is shown in tables 4 and 5, which show the eigenvalues greater than one [9] and the proportion of variance explained by first 8 factors. In order to provide additional examples, the PSE→ISR and LBN→ISR dyads were also analyzed.[10]

The results of this analysis are mixed. In all of the cases, the first factor explains over 50% of the variance, and in most instances around two-thirds of the variance. There is, clearly, a single dominant factor in the data.

However, in most instances there is also a second factor explaining a non-trivial additional 20% of the variance, and in the AFP ISR→LBN case there is even a third factor. Beyond this point the proportion of variance explained drops off quickly—the informal "scree test" for retained factors would place the cut-off at two—but this suggests that the unidimensionality could be rejected. [11] This has implications not only for the IRT models, but also the Goldstein, Vincent and Azar-Sloan scales.

Interpreting the *meaning* of factor scores is, as always, an occult art at best, but nonetheless I gave it a try, after first doing a `varimax` rotation on the factors to make them orthogonal. Eyeballing the plots, it appears that Factor 1 is, as expected, generally reflective of a cooperation-conflict dimension, or possibly just a dimension reflecting the presence or absence of conflict.

The second factor, on the other hand, is quite different for the two cases, which was evidenced both from visual examination and the more systematic comparison found in Table 6. This table rank-orders the CAMEO categories by the absolute value of the difference between the scoring coefficients on factors 1 and 2. Since factor 2 is orthogonal to factor 1, this should provide an indication of what types of behavior factor 2 is emphasizing that factor 1 was not.

What is most striking about Table 6 is the divergence between the rankings in the two cases. For example, the top three categories in the ISR→PSE case are ranked in the bottom four for the ISR→LBN; the top three for ISR→LBN are pretty much randomly scrambled, at positions 6, 10, and 17, for ISR→PSE. The two rankings basically have no relationship to each other.

This suggests that while there is a second factor in both cases, it varies by the case and essentially reflects whatever other major political trends are found in the sequence. For ISR→PSE, this is largely the distinction between the lull in violence (and attempts, however limited, at cooperation) during the Oslo period of the 1990s versus the high levels of violence following the outbreak of the *al-Aqsa intifada* in September 1990. Thus we see that the four

[9] One of the rules-of-thumb for inclusion of a factor, since a factor with an eigenvalue less than one explains less variance than one of the original variables.

[10] These data also include February, March and April of 2007.

[11] The Rizopoulos's package includes an option for testing a two-factor latent trait model and formally comparing its variance explained to that of alternative models; I will probably pursue this in the future.

Table 4: Eigenvalues > 1.0 and proportion of variance explained by first 8 factors, Israel↔Palestine

	ISR→PSE		PSE→ISR	
Eigenvalue	AFP	Reuters	AFP	Reuters
1	3.3359	2.2420	3.7219	2.7387
2	1.1476	0.8454	1.2381	0.8998
3	0.7521	0.6032	0.5783	0.6146
Proportion	AFP	Reuters	AFP	Reuters
1	0.6295	0.6632	0.6659	0.6049
2	0.2166	0.2501	0.2215	0.1988
3	0.1419	0.1784	0.1035	0.1358
4	0.0961	0.1553	0.0716	0.1052
5	0.0811	0.0843	0.0634	0.1046
6	0.0519	0.0603	0.0526	0.0892
7	0.0288	0.0343	0.0364	0.0592
8	0.0201	0.0110	0.0171	0.0306

Table 5: Eigenvalues > 1.0 and proportion of variance explained by first 8 factors, Israel↔Lebanon

	ISR→LBN		LBN→ISR	
Eigenvalue	AFP	Reuters	AFP	Reuters
1	7.8867	3.8894	5.5432	2.5191
2	1.6709	1.3944	1.3391	0.8064
3	1.1757	0.7559	0.5527	0.6560
Proportion	AFP	Reuters	AFP	Reuters
1	0.6940	0.6273	0.7286	0.5566
2	0.1470	0.2249	0.1760	0.1782
3	0.1035	0.1219	0.0726	0.1450
4	0.0492	0.0759	0.0631	0.1144
5	0.0275	0.0621	0.0495	0.1061
6	0.0249	0.0471	0.0289	0.0788
7	0.0128	0.0348	0.0205	0.0687
8	0.0103	0.0186	0.0157	0.0613

Table 6: CAMEO categories ranked by difference in scoring coefficients for Factors 1 and 2, AFP data

ISR→PSE				ISR→LBN			
CAMEO	Factor 1	Factor 2	Difference	CAMEO	Factor 1	Factor 2	Difference
03	0.313	-0.067	0.380	09	0.327	-0.092	0.420
04	0.319	-0.061	0.380	08	-0.100	0.305	0.405
19	-0.087	0.224	0.311	15	0.233	-0.097	0.331
18	-0.054	0.214	0.269	07	0.223	-0.080	0.303
13	-0.053	0.212	0.265	10	-0.081	0.150	0.231
08	0.158	-0.089	0.248	02	-0.076	0.153	0.230
16	0.175	-0.062	0.237	05	-0.078	0.135	0.214
01	0.077	0.273	0.196	14	-0.044	0.061	0.106
11	0.032	0.172	0.139	18	0.090	-0.012	0.103
15	-0.025	0.091	0.116	01	0.100	0.198	0.098
12	0.095	-0.013	0.109	13	0.044	0.125	0.081
07	-0.029	0.070	0.100	11	0.048	-0.032	0.080
02	0.022	0.120	0.097	16	0.006	0.067	0.061
05	0.109	0.030	0.079	12	-0.016	0.038	0.054
10	0.011	0.058	0.047	04	0.101	0.048	0.052
17	0.058	0.039	0.019	19	0.159	0.198	0.038
09	0.020	0.025	0.005	17	0.017	0.046	0.028
14	0.030	0.034	0.004	03	0.042	0.036	0.006

categories with the highest difference for ISR→PSE are two involving cooperative negotiation (03 and 04) and, negatively weighted, two involving use of force (18 and 19); the "threaten" category, also negatively weighted, comes in at position 5.

The ISR→LBN scores, in contrast, are much more difficult to interpret. One quite likely reason for this is that Israel is dealing with two groups in Lebanon: the Lebanese government (such as it was during much of this period) and the militarized non-state (or quasi-state) actor Hezbollah. Those interactions are quite different, with a mixture of cooperation and conflict vis a vis the government of Lebanon, and generally unremitting hostility towards Hezbollah. While all LBN actors were combined in this analysis, Hezbollah is coded separately in CAMEO—`LBNREBHEZ`—and it would be straightforward to pull this out and see what difference it makes.

Finally, as an exercise I correlated the factor scores produced for the AFP and Reuters series to see how well these would perform in terms of reconciling the two data sets. The correlation for the first factor is 0.62, which is somewhat better than the correlation in the IRT models, but still nowhere close to the correlation of the Goldstein series; the correlation for the second factor is only 0.46.

7 Comparison of Rasch and Goldstein scales

The final test involves whether the Rasch scores actually correspond to the Goldstein scores: the objective here is to see whether a scale can be inductively created for CAMEO data that provides a time series similar to that provided by the expert-derived Goldstein scale for WEIS data. The Rasch model is used here since it is the simplest of the IRT models, and the two-parameter models correlate quite highly with the Rasch values.

As reported in Tables 2 and 3, we already know the basic correlations: for the Palestinian case, -0.46 for AFP and -0.43 for Reuters; for the Lebanon case -0.68 for AFP and -0.62 for Reuters. (The negative correlation is due to the Goldstein scale using negative values for conflict, while the Rasch intensity dimension increases positively with greater activity.) These numbers aren't bad—and are significant at the $p < 0.001$ level—but still aren't great.

A visual examination of the plots of the Rasch scores showed that one of the problems is that these fluctuate considerably more rapidly than the Goldstein scores, but appeared to be generally fluctuating around the Goldstein values. Consequently I undertook the following "mild" data-fitting exercise to see whether the two AFP series could be brought into closer alignment

- Take a 5-month centered moving average of each series
- Regress the Rasch score on the Goldstein score and use the coefficients of that linear regression to make the Rasch score comparable in magnitude to the Goldstein score[12]

The results of this exercise are shown in Figures 10 and 11; the correlation is 0.57 for Palestine and 0.82 for Lebanon.[13] There is a systematic divergence in the Palestine data,

[12] The rescaling was $new = -132 * old - 62.3$ for Palestine and $new = -33 * old - 26$ for Lebanon.

[13] For purposes of comparison of the effects of the moving average, the the correlations for the moving averages of the AFP and Reuters series are 0.64 for Palestine and 0.71 for Lebanon.

with the Rasch score almost uniformly lower prior to the outbreak of the *al-Aqsa intifada* and higher afterwards. Since the key difference between these two periods is the incidence of armed violence, the distinctions may be largely attributable to only those categories. The Lebanon series lines up very closely, albeit again with an exception in the spike of violence in the summer of 2006, where the Goldstein score is almost twice the magnitude of the Rasch score.

While these correlations are perhaps not as high as one would like, the convergence appears fairly remarkable given that the Rasch scaling is a fully-automated, purely inductive method that involved no *a priori* information concerning the Goldstein scale. Depending on ones perspective, the Rasch method either duplicated the expertise of Goldstein's panel, or Goldstein's panel successfully groked[14] the primary underlying latent dimension of event data. Either way, these results suggest that Rasch scaling could be used to develop event scales in an objective, replicable fashion rather than developing these using expert panels.

[14] If I have to explain this concept, you won't understand it. Google it...

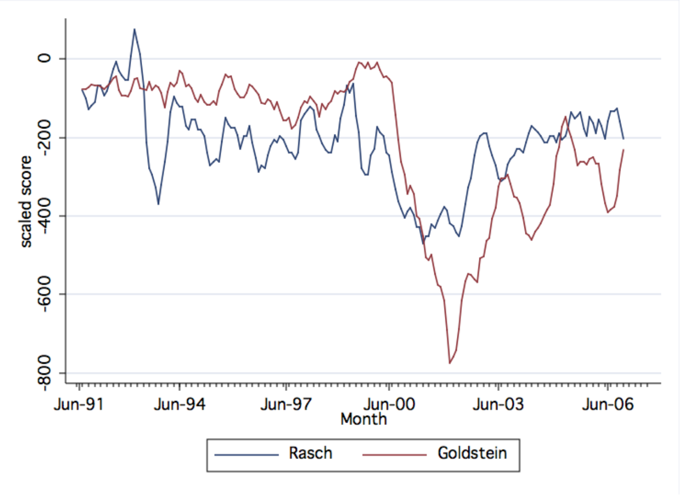

Figure 10: Monthly rescaled Rasch scores and Goldstein scores for AFP having Israel as actor and Palestinians as target, 5-month moving average

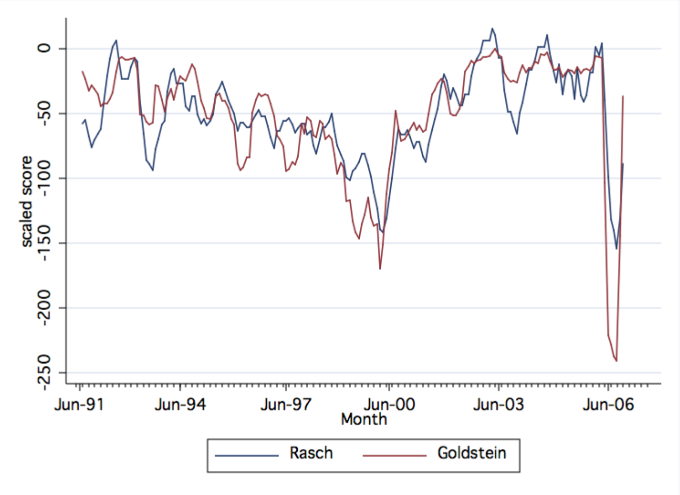

Figure 11: Monthly rescaled Rasch scores and Goldstein scores for AFP having Israel as actor and Lebanon as target, 5-month moving average

8 Conclusion

The results of this exercise have been mixed. To start with the most positive result, the basic concept that IRT can be used to generate credible scales for event data would seem to have been validated, albeit more for the Lebanon case than the Palestinian. In the Lebanese case, the correlation between the moving averages of the Rasch scores and Goldstein scores (0.82) is almost as high as the correlation between the Goldstein scores computed from AFP and Reuters sources, despite the very substantial reduction of information contained in the data used to compute the Rasch scores. The Palestinian case aligns less well—a Rasch/Goldstein correlation of 0.57 and AFP/Reuters correlation of 0.80—but still relatively high.

These results are particularly impressive given the paucity of information that went into computing the IRT scores: the original data were standardized to a point where all absolute frequency information was discarded, and no *a priori* information was provided about the relative importance of the categories. The calculations were performed in a minute or two, versus the extended (and in some instances unsuccessful) scaling exercises that try to elicit information from subject matter experts. This method would, therefore, appear appropriate for the automated development of scales for new coding systems.

The technique was less useful for the purpose I originally proposed: reconciling multiple news sources. It is credible, with inter-source correlations for the Rasch model of 0.49 for ISR→PSE and 0.62 for ISR→LBN, but both scores are substantially lower than the correlations of the Goldstein scale—0.80 and 0.90 respectively—and the additional parameters provided by estimating a Birnbaum model do not really improve on the simple Rasch model.

To a certain extent, the inability to improve on the Goldstein scale is due to the fact that the AFP and Reuters scores diverge less on the new series based on Reuters stories downloaded from the Factiva data service than did some earlier series we produced based on stories downloaded from Reuters own [awful] data services, particularly during the period of Reuters's institutional near-death-experience in the late 1990s. With the new data, the inter-source comparison problem is less than originally anticipated, and it would be difficult to improve on those scores in any circumstances.

However, a more fundamental problem lies in the inductive character of the estimation, and the assumption that common events are "easy" and hence less important. The point where this makes the Goldstein and IRT approaches diverge substantially is the treatment of uses of force. For these dyads, particularly ISR↔PAL during the 2000s and ISR↔LBN in the 1990s, violence is an all too common event, and consequently received a low weight in terms of "difficulty" on the latent trait. This is not due to the absolute frequency, since the mean-based dichotomous coding scheme effectively only provides information on the skew of the data, but rather the fact that in these two intensely, if ineffectually, mediated conflicts, months in which a lot of things are happening are likely to be months when violence is happening. The "guessing" estimates, for example, generally had the c_i estimates for the violence category 19 at about the same level at category 01, the innocuous catch-all "Make public statement." The expert weighting of the Goldstein scale, in contrast, takes into account the fact that *globally*, force events are relatively rare, and consequently receive a high weight.

The most straightforward inductive solution to this problem would be to re-estimate values from a global event data set such as the original WEIS or the more recent VRA

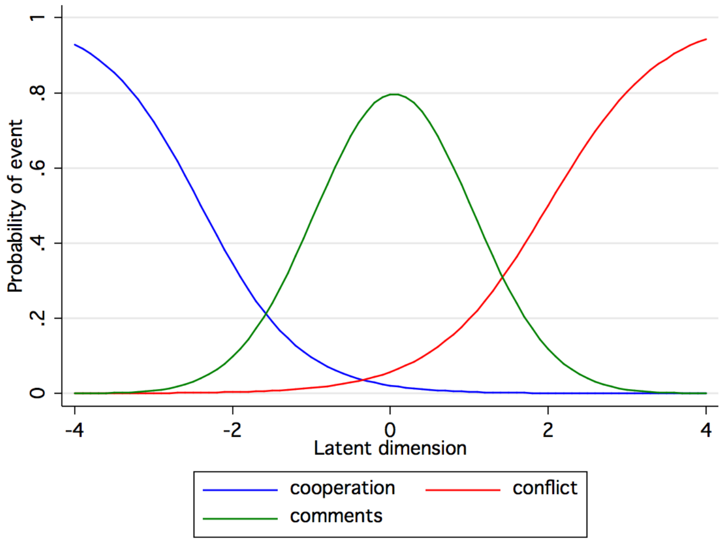

Figure 12: Hypothetical mix of probability curves for different types of events

set. Unfortunately we can't do this with CAMEO since the KEDS project focuses almost exclusively on protracted conflict situations, and virtually all of our data sets contain levels of violence that are disproportionate to those in a randomly selected dyads.

A second problem may lie in the differences between the assumptions of the relation between the latent trait and the probability of various events. IRT assumes the structure of a Guttman scale, where for any pair of items x and y,

$$\ell(x) > \ell(y) \Rightarrow Pr(x) < Pr(y) \tag{3}$$

where $\ell()$ is the location of the item on the latent trait. This is not necessarily the case for events: in particular, as conflict events become more likely, cooperative events might be expected to become less likely, with neutral events somewhere in the middle. Consequently the actual relationship may look more like Figure 12.

In fact, Figure 12 seems so much more plausible than the Guttman-scaled assumption that one might question why this analysis worked even to the extent that it did. In all likelihood, this is due to two characteristics of the conflict. First, there is relatively little actual cooperation in these dyads—even the Oslo "peace process" was at best more of an extended truce, and active cooperation broke down quickly after the assassination of Rabin in the fall of 1995—so the latent dimension is more along the lines of "no conflict" to "high conflict" rather than "cooperation" to "conflict." Second, both conflicts experience a great deal of external mediation, so as violence increases, it is the case that various forms of consultation and commentary will also increase (more so with ISR→PAL than ISR $rightarrow$LBN).

There are two possible solutions to this. The first—and more straightforward—is simply to estimate the cooperative and conflictual events separately, and then combine these scales, for example, by assigning positive numbers to the latent trait of the cooperative events and negative numbers to those of the conflictual events, an approach similar to that of both the Azar-Sloan and Goldstein scales. There is a fairly clear distinction between conflictual and cooperative events in most event coding systems and within each of those types, the Guttman assumption is more reasonable: For example a dyad that is experiencing uses of force will almost certainly also be experiencing demands, threats and protests; a dyad experiencing an exchange of aid will also experience consultations and diplomatic cooperation.

Another possibility would be to see if it is possible to modify the IRT estimators to handle multiple curves. This would seem particularly straightforward for the downward-sloping logistic—the "cooperation" curve in Figure 12—since this is identical to an upward-sloping logistic except for a negative value of α.[15] Handling the intermediate "comment" case might be more challenging, though perhaps a parameterized family of curves exists that could do this.

A final possible extension of this approach that I may pursue in the future involves retaining greater information in the reduction of the events to scores that can be analyzed using IRT. The current method of reducing the event data to the dichotomous variables required by the IRT method is, as noted at numerous points, quite extreme. In particular, by adjusting every category relative to its own mean, it removes the inter-category variation in frequency, as well as reducing the variation of that frequency. This could be adjusted by scoring the months against absolute counts (possibly logged); additional variation could be provided by using a polytomous scoring system, perhaps against quartiles or quintiles.

[15] The estimates of α_i in the two-parameter discrimination models were all positive; at present I don't know whether this is due to a constraint in the estimation procedure or a characteristic of the data.

9 Appendix: CAMEO Code Summary, Version 0.9B2

01: MAKE PUBLIC STATEMENT
02: APPEAL
03: EXPRESS INTENT TO COOPERATE
04: CONSULT
05: ENGAGE IN DIPLOMATIC COOPERATION
06: ENGAGE IN MATERIAL COOPERATION
07: PROVIDE AID
08: YIELD
09: INVESTIGATE
10: DEMAND
11: DISAPPROVE
12: REJECT
13: THREATEN
14: PROTEST
15: EXHIBIT FORCE POSTURE
16: REDUCE RELATIONS
17: COERCE
18: ASSAULT
19: FIGHT
20: ENGAGE IN UNCONVENTIONAL MASS VIOLENCE

References

[1] Azar, Edward E. 1980. "The Conflict and Peace Data Bank (COPDAB) Project." *Journal of Conflict Resolution* 24:143-152.

[2] Azar, Edward E. 1982. *The Codebook of the Conflict and Peace Data Bank (COPDAB).* College Park, MD: Center for International Development, University of Maryland.

[3] Azar, Edward E., and Thomas Sloan. 1975. *Dimensions of Interaction.* Pittsburgh: University Center for International Studies, University of Pittsburgh

[4] Baker, Frank B. and Seock-Ho Kim. 2004. *Item Response Theory.* New York: Marcel-Dekker.

[5] Birnbaum, A. 1968. "Some Latent Trait Models and Their Use in Inferring an Examinee's Ability." In F. Lord and M. Novick, eds. *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley

[6] Bond, Doug, Joe Bond, Churl Oh, J. Craig Jenkins and Charles Lewis Taylor. 2003. "Integrated Data for Events Analysis (IDEA): An Event Typology for Automated Events Data Development" *Journal of Peace Research* 40,6: 733745

[7] Bond, Trevor G. and Christin M. Fox. 2001. *Applying the Rasch Model.* Mahwah, NJ: Lawrence Erlbaum.

[8] Embretson, Susan E. and Steven P. Reise. 2001. *Item Response Theory for Psychologists.* Mahwah, NJ: Lawrence Erlbaum.

[9] Deborah J. Gerner, Philip A. Schrodt, Ömür Yilmaz and Rajaa Abu-Jabr. 2002."Conflict and Mediation Event Observations (CAMEO): A New Event Data Framework for the Analysis of Foreign Policy Interactions." Paper presented at the International Studies Association, New Orleans, March 2002 and American Political Science Association, Boston, August 2002.

[10] Deborah J. Gerner, Philip A. Schrodt, Ömür Yilmaz and Dennis Hermrick. 2005. "The CAMEO (Conflict and Mediation Event Observations) Actor Coding Framework." Paper presented at the American Political Science Association, Washington, September 2005.

[11] Deborah J. Gerner, Philip A. Schrodt and Ömür Yilmaz. 2007. *Conflict and Mediation Event Observations (CAMEO) Codebook.* Manuscript, `http://web.ku.edu/keds/data.dir/cameo.html`

[12] Goldstein, Joshua S. 1992. "A Conflict-Cooperation Scale for WEIS Events Data." *Journal of Conflict Resolution* 36: 369-385.

[13] International Studies Quarterly. 1983. "Symposium: Events Data Collections." *International Studies Quarterly* 27.

[14] Leng, Russell J. 1987. *Behavioral Correlates of War, 1816-1975.* (ICPSR 8606). Ann Arbor: Inter-University Consortium for Political and Social Research.

[15] King, Gary and Will Lowe. 2003. "An Automated Information Extraction Tool for International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design." *International Organization* 57,3: 617-642.

[16] McClelland, Charles A. 1976. *World Event/Interaction Survey Codebook.* (ICPSR 5211). Ann Arbor: Inter-University Consortium for Political and Social Research.

[17] McClelland, Charles A. 1983. "Let the User Beware." *International Studies Quarterly* 27,2:169-177

[18] Rizopoulos, Dimitris. 2006. "ltm: An R Package for Latent Variable Modeling and Item Response Theory Analyses." *Journal of Statistical Software* 17, 5. http://www.jstatsoft.org/

[19] Schrodt, Philip A. and Deborah J. Gerner. 1994. "Validity Assessment of a Machine-Coded Event Data Set for the Middle East, 1982-1992." *American Journal of Political Science* 38:825-854.

[20] Shellman, Stephen M. 2004. "Measuring the Intensity of Intranational Political Interactions Event Data: Two Interval-Like Scales." *International Interactions* 30,2: 109-141.

[21] Vincent, Jack E. 1979. *Project Theory: Interpretations and Policy Relevance.* Washington: University Press of America.