

Reply to Spagat

March 31, 2016

A researcher named Michael Spagat has posted comments online making false accusations against our company and an unpublished paper alleging potential falsification in data produced by two of our field work providers in Iraq, D3 Systems Inc. and KA Research Ltd. We investigated these claims in 2011 and found them to be baseless. Our analysis follows.

--

Review of the Koczela/Spagat paper, August 2011

We have received a draft of an unpublished academic paper by Steven Koczela and Michael Spagat that claims, through statistical analysis, to find “evidence of potential fabrication” in public opinion surveys in Iraq by KA Research and D3 Systems. The authors call on other KA/D3 clients to examine their own Iraq data for similar patterns. We have done so.

This memo details our review of the four Iraq polls produced by KA and D3 for ABC News and its media partners from 2007 through 2009. In replicating the analysis conducted by Koczela and Spagat, we find most of the same patterns they report using other KA/D3 Iraq datasets. In extending this analysis, however, we find that these patterns are an artifact of the authors’ groupings of field work supervisors. Across all supervisors individually we see no patterns that would support a charge of fabrication.

An allegation of fabrication of data is the most serious charge a researcher can make. While we can speak only to our own ABC data from Iraq, our review indicates that the analysis by Koczela and Spagat, while compelling on first read, is in fact selectively and misleadingly presented. We encourage D3/KA and its other Iraq clients not only to replicate the Koczela/Spagat analysis, but to extend it as we describe below. We also welcome suggestions for additional analysis.

Beyond our statistical analysis, we do have some documentary evidence in hand. KA/D3’s work for us in Iraq included delivery of interviewer and supervisor journals describing their field work experiences, and photos of field work as it occurred. Our review finds that we have both journals and photos of field work from the areas where Koczela and Spagat suggest that field work did not occur.

While we focus below on the statistical evidence, we note that Koczela and Spagat are silent on an important related issue, the mechanism of the fraud they allege. The falsification they suggest would have required a remarkable conspiracy – collusion across seven supervisors, coordinating faked questionnaires with the same pattern of responses, with either ghost interviewers or further

Review of ABC News Iraq data – August 2011

collusion by interviewers and deception of other staff, and further producing faked field journals and staged photos. It's difficult to comprehend how or why this would have been undertaken en masse; collusion would have increased, not reduced, any individual supervisor's odds of apprehension.

Most fundamentally, however, our data simply do not support the allegation.

Analysis

Koczela and Spagat charge that patterns in KA/D3 Iraq data produced under seven specific supervisors (whom they call “focal” supervisors) are inexplicably different from the data produced under the direction of the 16 to 20 other supervisors on any given poll. They conclude that these patterns constitute potential evidence of fabrication.

We have examined our own Iraq data for evidence of four of the five types of patterning identified by the authors – partial distributions of substantive responses, empty categories of “don't know” and “refused” responses, “implausible correlations” and restricted ranges in interval scale variables. Each is evaluated in detail below. We cannot speak to their fifth point, relating to household TV viewership, as we lack such data.

Each ABC dataset includes work by at least some of the supervisors identified as problematic by Koczela and Spagat, enabling us to replicate as well as to extend their analysis. Focal interviewers accounted for 31 to 36 percent of total interviews in the ABC datasets, and nearly all of the interviews in Anbar, Baghdad and Diyala.

Supervisor number:	ABC 3/07 (#1033)	ABC 8/07 (#1043)	ABC 2/08 (#1060)	ABC 2/09 (#1087)
36		Diyala	Diyala	Diyala
38	Diyala			
43	Baghdad	Baghdad	Baghdad	Baghdad
44	Anbar	Anbar	Anbar	Anbar
47	Baghdad	Baghdad	Baghdad	Baghdad
93		Baghdad	Baghdad	Baghdad
94		Baghdad	Baghdad	Baghdad
# of other supervisors	16	17	18	20
Focal supervisor case count	695	795	795	719
Other supervisor case count	1517	1417	1433	1509
Total sample size	2212	2212	2228	2228

While, again, we find no evidence of fabrication, there are aspects of the analysis that merit further examination. Partial distributions and DK/Refs may speak to inconsistent training and field work procedures, although other factors – translation, sociocultural norms and the role of

armed conflict in intensifying or polarizing attitudes – also may be at play. The authors correctly point out that quality-control efforts customarily are focused at the interviewer level; a review of these efforts in general, and specifically in terms of oversight of field work supervisors, is advisable. Finally, the allegations in the paper underscore the need for the fullest possible documentary evidence of field work procedures.

a. Partial Distributions of Categorical Variables

Paper Summary:

A “partial distribution” means a question in which some available response option – e.g. “strongly agree,” or “very satisfied” – is not selected by any respondent. Koczela and Spagat find a high number of questions with partial distributions among the seven “focal” supervisors they have selected for comparative analysis. Specifically, in the focal data they reviewed, 23 to 53 percent of the questions have partial distributions, while among the non-focal supervisors this never happens. They say, “complete unanimity on one category out of six offered categories is almost unheard of,” and “the empty response categories that are frequently present among the focal supervisors, and conspicuously absent among the rest, are inconsistent with ... known realities of survey research.”

Koczela and Spagat excluded questions they regarded as most likely to have strong regional influence, a debatable approach because it’s judgmental. Also, their calculation of the probability that the differences in distributions could have happened by chance (essentially nil) is based on the assumption that the non-focal supervisors’ data is the “true” distribution, which is problematic given regional differences.

Regardless of these limitations, evidence of these differences, as presented, seems strong, in the absence of deeper evaluation. We find, however, that the patterns dissolve when we look at all supervisors on an individual level, rather than aggregating them as Koczela and Spagat have done. The authors’ claim that “empty response categories are...conspicuously absent among [the other supervisors]” is entirely misleading (at least in terms of our ABC data), because it’s the case only when those supervisors are collapsed into one large group. In fact, in ABC data, not one supervisor alone has completely distributed data.

ABC Data:

Our initial goal was to see whether the patterns identified by Koczela and Spagat as problematic were present in our own data. To do so, we compared the number of questions with partial distributions among these interviewers vs. the rest, by creating a binary variable differentiating “focal” and “other” supervisors and running crosstabs on the unweighted frequencies for each attitudinal question (demographics were analyzed separately). We did not attempt to eliminate any questions that might have a regional influence.

As summarized in the table below, we found that 44, 49, 40, and 40 percent of the questions in the focal supervisors’ data had at least one substantive (non-DK/Ref) response category

completely empty (i.e., had partial distributions). In the non-focal supervisors’ data, aggregated, there were no questions with partial distributions.

Also replicating Koczela and Spagat, we found some differences in demographic data between the two supervisor groups. In two of our datasets, no respondents in the focal data reported being separated, divorced or widowed. In all four polls, all respondents in the focal supervisor data indicated having a shortwave radio. In two polls, no respondents in the focal data reported being unemployed and not looking for a job.

	Dataset (Count of categorical variables with partial distributions among the 7 focal supervisors)								
	BBG 12/05	PIPA 1/06	PIPA 9/06	ABC 3/07	ABC 8/07	BBG 9/07	ABC 2/08	BBG 11/08	ABC 2/09
	(167)	(57)	(60)	(113)	(82)	(93)	(102)	(83)	(114)
Count	54	23	14	50	40	49	41	31	46
%	32%	40%	23%	44%	49%	53%	40%	37%	40%

The question, however, was whether grouping supervisors into “focal” and “other” groups, with the latter group being much larger, might obscure how common it was for individual supervisors to have partially distributed data.

To test this hypothesis we examined 120 questions (this included all attitudinal and demographic questions, but excluded open-ends and those asked only of subsamples) in the ABC #1087 data by individual supervisors, and tracked how many supervisors had partial distributions on each question. (We focused for this purpose on empty substantive response categories, leaving aside DK and Ref for separate analysis, below.) The findings show that partially distributed data were not, in fact, “almost unheard of,” but instead rather common across almost all of the questions.

On 95 questions (79 percent) at least one supervisor had partially distributed data, and on three questions all 25 supervisors had partially distributed data. The average number of supervisors to have partially distributed data on a question was eight, and the average number of partially distributed questions per supervisor was 38.

While the focal supervisors Koczela and Spagat identified had partially distributed data on more questions than most of the other supervisors (between 55 and 67 questions), there were partial distributions for the other supervisors as well (between 10 and 65 questions). It’s plausible, moreover, that opinions in the hotspots of Baghdad, Anbar and Diyala may have shown greater unanimity of responses because of the extreme conditions there.

This reasoning would have been weaker had there been no partial distributions anywhere else. But there were. Some specific examples from the ABC #1087 data follow:

- Sixteen of the 25 supervisors had at least one empty response category in their data on Q2 (expectations for their life a year from now), suggesting (given that supervisors vary by

region) that expectations for the future tend to depend on current experiences in one's region. "Much worse" was not selected in 16 of the 25 supervisors' data, and "much better" was not selected in data from seven of the supervisors.

- On Q6, not one respondent in the focal supervisors' data said they thought things in Iraq would be "much worse" a year from now. That seemed odd – until we found that in fact no one selected "much worse" in data from 18 supervisors, out of 25.
- On Q14, the grouped analysis shows that only focal supervisors have partially distributed data, with not one person selecting that they thought Iraq should be "a country divided into separate independent states." However, in addition to the six focal supervisors in this poll, 12 other supervisors also had data where not one person selected that option, and one supervisor in the "other" group had no one select another option.
- There were a number of questions where not one supervisor had partially distributed data, but there are no questions where the six focal supervisors were the only supervisors to have partially distributed data. This suggests that if indeed the focal supervisors had fabricated their data, they somehow knew to make sure to utilize all response options in some cases (when every other supervisor did), but not in others (when at least one other supervisor had partially distributed data) – or else it is a remarkable coincidence. The far more plausible explanation is that responses to some of the questions depended greatly on local experiences (which in Baghdad, Diyala and Anbar might be particularly unifying), whereas other questions tended to tap more personally held beliefs.
- Moreover, there are 21 questions where none of the six focal supervisors had partially distributed data, but one or more of the other supervisors did (Q9h, Q27a, Q27b, Q27d, Q27e, Q27f, Q27g, Q27h, Q27i, Q27j, Q29, Q31, Q34c, Q36, Q45b, Q45d, Q45e, Q46, D15c, D15i, D15l). Collapsing the other supervisors together obscures these cases.
- Finally, it should be noted that while the focal supervisors were responsible for all of the data collected in Baghdad and Diyala, the focal supervisor collected just half of the data in Anbar for this poll. The other half was collected by a supervisor who was not implicated in the Koczela/Spagat paper as being suspected of fabrication. Yet the partial distribution rates for the two Anbar supervisors were identical, with each having at least one empty response category on 65 of the 120 questions. This further suggests that elevated partial distribution rates among the focal supervisors are a function of regional differences, not evidence of collusive fabrication.

In summary, it is not surprising that grouping together 16 to 20 supervisors reduces the likelihood of showing empty cell counts compared to a grouping of just six supervisors. But doing so when many of the "other" supervisors also show partially distributed data is misleading. This approach suggests that the focal supervisors are the only supervisors with partially distributed data, which is not at all the case.

b. Empty Categories of "Don't know" and "Refused"

Paper Summary:

Koczela and Spagat report that on many questions no respondents in the focal supervisors’ data answered don’t know, or refused to answer, while data from the other supervisors had positive numbers in both categories in the same questions. Specifically, for 30 percent to 72 percent of the questions these focal supervisors had zero counts for don’t know and refused, whereas this never happened for the other supervisors on those questions. The authors suggest that data fabricators may attempt to make their data look “clean” in order to avoid red flags in quality control, and suggest this may include a tendency to avoid DK and Ref options.

Our evaluation of the DK and Ref patterns identified by Koczela and Spagat again finds them to be an artifact of aggregation.

ABC Data:

As with partial responses, unweighted crosstabs of our results show the same pattern described by Koczela and Spagat across all four polls. From 56 percent to 91 percent of the attitudinal questions have empty DK or refusal counts in the focal sample, but not in the other sample. (In other words, only questions where someone in the other sample responded with a DK or Ref but no one in the focal sample did were counted.) It is also the case that for ABC survey #1043 (the one examined in detail here), there were no Refs across all 82 questions examined for the focal supervisors group, and only one item with non-zero DKs. In contrast, the other supervisors group had no refusals on 41 percent of the questions examined and empty DK cells for only 13 percent of the questions.

Looking just at DK responses, in ABC #1043 there are 70 questions out of 82 in which the focal supervisors group has a zero DK count but the other supervisors do not. (There are 11 questions in which both the focal and other groups have zero DK counts, and one in which both groups have a non-zero count on DK.) For refusals, the focal group has a zero count on all 82 questions, but there also are 34 questions where the non-focal group has a zero count.

	Dataset (Count of questions with empty DK or Ref counts in the focal supervisor data but non-zero DK and Ref counts in the non-focal data)								
	BBG 12/05	PIPA 1/06	PIPA 9/06	ABC 3/07	ABC 8/07	BBG 9/07	ABC 2/08	BBG 11/08	ABC 2/09
	(167)	(57)	(60)	(113)	(82)	(93)	(102)	(83)	(114)
Count	50	41	37	63	48	42	93	45	64
%	30%	72%	62%	56%	59%	45%	91%	54%	56%

In many cases, however, while there are DK or Ref answers in the non-focal group, there are very few of them – even just one or two. That means that zero DK/Ref responses are found among many of the individual non-focal supervisors’ data as well as among focal supervisors, a phenomenon concealed by aggregation.

Review of ABC News Iraq data – August 2011

Our first question was the magnitude of the difference across items where focal supervisors had a zero count and other supervisors had a non-zero count. We measured this by sorting the questions with empty DK/Ref counts in the focal group into those in which 5 or fewer, 6-10, or 11+ respondents answered DK or Ref in the non-focal group.

	DK (n=70)		Ref (n=48)	
	Count	%	Count	%
5 or less	32	44%	29	60%
6 to 10	18	26%	10	21%
11 plus	21	30%	9	19%

As the table shows, on many questions the number of people in the other supervisors' data who gave DK or Ref responses was quite small. In 44 percent of the cases for DKs and 60 percent of the cases for Refs five or fewer people gave these responses, not exactly a big leap to zero.

The key analysis, however, is whether the non-zero counts hold when the non-focal supervisors group is disaggregated. The table above suggests that it cannot, because in most cases there are fewer individuals giving DK/Ref responses in the non-focal group than there are supervisors, meaning that at least some of the supervisors must have zero counts. Indeed, we find that zero DKs and Refs are far more common for individual non-focal supervisors than Koczela and Spagat's aggregated comparison suggests. Supervisors may have been zealous in having interviewers encourage respondents to give an answer, not DK or Ref. But if missing DK and Ref responses are not isolated to focal supervisors, the case for fabrication does not stand.

Don't Knows

Of the 17 non-focal supervisors, one has zero DKs across all 82 items, the same as the focal supervisors. Seven non-focal supervisors have zero DKs on 95 percent of the items or more. The other nine non-focal supervisors have zero DKs on 12 to 59 percent of the items.

Five supervisors in particular have especially high DK rates, accounting for the lion's share of DKs among the non-focal supervisors. The other 12 non-focal supervisors have zero DKs on 75 percent or more of the items.

	Supervisor																	
	2	6	16	23	25	29	32	40	46	50	54	58	63	64	77	80	91	Total
DKs	2	6	16	23	25	29	32	40	46	50	54	58	63	64	77	80	91	82
Items	12	1	10	3	14	4	3	26	15	26	4	2	48	3	0	22	25	82
%	15%	1%	12%	4%	17%	5%	4%	32%	18%	32%	5%	2%	59%	4%	0%	27%	30%	

Another approach is to add the total number of DKs across items, rather than counting the number of items on which there was at least one DK. As the table below shows, one supervisor (Supervisor 50) accounts for one-quarter of all the DKs among the non-focal supervisors. Supervisors 80, 40, and 63 account for a further 14, 13, and 11 percent, respectively. These four supervisors account for 63 percent of DKs recorded among the 17 non-focal supervisors. The

Review of ABC News Iraq data – August 2011

rest have much lower total DK counts, including 11 supervisors who account for 5 percent of the DKs or fewer each.

	Supervisor																	
DKs	2	6	16	23	25	29	32	40	46	50	54	58	63	64	77	80	91	Total
Count	79	2	24	3	25	45	3	115	51	214	38	2	99	4	0	126	44	878
%	9%	<1%	3%	<1%	3%	5%	<1%	13%	6%	24%	4%	<1%	11%	<1%	0%	14%	5%	

Using one specific illustration, Q14f has 163 DKs in the non-focal group compared to zero DKs in the focal group. However, closer inspection reveals that 109 of these DKs are found in Supervisor 50 and another 42 in Supervisor 29. Most of the other non-focal supervisors (11 out of the remaining 15) have zero counts, the same pattern as the focal supervisors.

Refusals

Our analysis of refusal rates has similar results. In this case, non-focal supervisors look even more like focal supervisors once they are disaggregated. Five non-focal supervisors did not have any refusals recorded, nearly as many as the six focal supervisors who did. An additional four supervisors have Ref rates of 5 percent or lower (excluding zero). Stated in the other direction, more than half the non-focal supervisors have zero refusals on 95 to 100 percent of the items.

As with the DK analysis, a small number of supervisors have higher Ref rates across the items. Supervisor 40 recorded at least one refusal on 16 percent of the items, as did Supervisor 63 on 22 percent and Supervisor 29 on 32 percent of the questions. The rest of the supervisors had very few or no refusals.

	Supervisor																	
Refs	2	6	16	23	25	29	32	40	46	50	54	58	63	64	77	80	91	Total
Items	0	2	7	0	3	0	3	26	4	13	7	0	18	3	0	9	3	82
%	0%	2%	9%	0%	4%	32%	5%	16%	9%	0%	9%	0%	22%	4%	0%	11%	4%	

Also as with DKs, adding total refusals across each supervisor provides further evidence against the charge of falsification. Two supervisors (40 and 54) account for more than half of the total refusal count among non-focal supervisors. As noted above, there are five supervisors who have zero Refs across all items, and another six who account for 1 to 5 percent of the Refs. In terms of total refusal count, the individual non-focal supervisors do not look markedly different from focal supervisors. There are many cases of low refusal and even zero refusal rates in the non-focal group. The bulk of the refusals are concentrated among a small group of supervisors, leaving most of the rest at or near zero.

	Supervisor																	
Refs	2	6	16	23	25	29	32	40	46	50	54	58	63	64	77	80	91	Total
Count	0	2	17	0	6	0	4	91	4	39	62	0	28	3	0	30	3	289

%	0%	1%	6%	0%	2%	0%	1%	31%	1%	13%	21%	0%	10%	1%	0%	10%	1%	
---	----	----	----	----	----	----	----	-----	----	-----	-----	----	-----	----	----	-----	----	--

In sum, both total count and item counts for the non-focal supervisors lead to very similar conclusions as the partial distribution analysis. The non-zero DK and ref rates that set apart the non-focal and focal supervisors fall apart when the data are analyzed by individual supervisor.

c. “Improbable correlations”

Paper Summary:

Koczela and Spagat provide correlations between related questions within each of the datasets, using only Sunni Arab respondents to control for ethnic differences. They find that the relationships between questions show the expected positive correlation in the other supervisors’ data but essentially zero correlations in the focal supervisors’ data.

This aspect of the paper is not well-documented. The authors only provide examples of one question in each of the datasets where this occurs, raising questions of cherry-picking. Moreover, some of the correlations between similar questions in the other supervisors’ data are relatively weak. However, zero correlations are rare to find, and one would expect to see some sort of positive relationship between the questions they selected.

Examining expected correlations is a reasonable way to search for evidence of data fabrication; it’s very difficult for a fabricator to anticipate relationships among variables and fake data accordingly. We find, however, that the lack of correlations of the type that Koczela and Spagat document appears again to be an artifact of their groupings of supervisors. (We also note that we have examined many more correlations, 96 in total, than Koczela and Spagat report.)

One key point is that a correlation coefficient assesses the amount of variance in two variables that is shared. Partially distributed data have less variance to be explained, therefore the chance of finding a significant correlation is greatly reduced. To the extent that partial distributions are more common in the focal group, correlations will be attenuated. Indeed, for eight question pairs, correlation coefficients could not be computed at all for the focal supervisor data because all respondents on one of the questions in the pair gave the same response (meaning there was no variance to explain).

In addition, the focal group consists of data from three provinces, whereas the other group contains data from 15 provinces. If one province in each group has a different correlation than the others, it’s going to have a much larger negative impact on the overall correlation in the smaller (focal) group than it would have on the larger (other) group.

ABC Data:

We examined ABC #1043 in two ways. First, following Koczela and Spagat, we looked at the correlations among conceptually related questions separately for focal and other supervisor groups. Second, we examined these same correlations for each of the provinces included in the

focal group (Anbar, Baghdad, and Diyala) vs. provinces with adequate unweighted samples sizes in the other group (i.e., with over 100 interviews apiece – Basrah, Tamim, Babylon, Dhi Qar, Neneveh, Suleymaniyah and Irbil), to see whether any zero, near-zero or negative correlations we found might be due to differing patterns at the provincial level.

It was difficult to choose questions in the ABC dataset that would obviously highly correlate (Koczela and Spagat had the fortune of finding two cases of nearly identical questions). However, we tested the correlations between questions on local conditions, the effects of U.S. forces, security, police, and teachers, which should have some kind of positive relationship.

Comparing focal and other supervisors there are a number of near-zero and negative correlations for the focal group where one would expect a positive relationship. However the pattern is not as consistent as Koczela and Spagat imply with their five examples. Out of the 88 possible question pairs examined, 36 pairs (40 percent) displayed near-zero correlations (+/- .10, one -.13) among the focal supervisors, while there was only one near-zero correlation among the 96 pairs for the other supervisors.

Out of the 88 pairings where it is possible to compare the magnitude of the correlations between the focal and other groups, there are only 14 cases where the correlations are not statistically different. Of those that are significantly different, 48 correlations (65 percent) are significantly stronger (in the expected positive direction) for the other group. But contrary to Koczela and Spagat, 26 (35 percent) are in fact stronger for the focal group.

Moreover, when cases in which the focal group had a partial distribution on at least one of the questions in the pair are removed, 26 of the 39 correlations (67 percent) are stronger for the focal group, while 13 (33 percent) are stronger for the other group. Therefore there is not a consistent pattern of the other group showing a stronger expected positive relationship, especially when we account for partial distributions.

The most consistent pattern of near-zero correlations in the focal group occurs for pairs of questions where there is reduced variability due to the partial distributions. Out of the 36 near-zero or negative correlations in the focal supervisors group, none involves pairings where both questions have full distributions. Thirty-one of the 36 pairings have partial distributions on both variables (86 percent of the near-zero or negative correlations) and 5 pairings (14 percent) have one variable with a partial distribution. Lack of correlation where one would expect to see one is entirely plausible in these cases simply because of the lack of full distributions.

Further, just as it did with the above analyses, disaggregating these data reveals a different story from Koczela and Spagat's. The provincial analysis described below reveals that low, near-zero, and negative correlations among the pairs of questions we tested are not only present within focal group provinces, but also within other provinces as well.

	Basrah	Tamim	Babylon	Dhi Qar	Neneveh	Suleymaniyah	Irbil
Negative or near-zero rs	1 (1%)	14 (15%)	3 (3%)	6 (6%)	26 (27%)	39 (41%)	53 (55%)

There also is a great deal of variability just between the three provinces where the focal supervisor interviews are concentrated. Anbar has a high rate of near-zero or negative correlations, accounting for nearly all of the pairings. Baghdad has a much lower rate (41 percent) and Diyala has an even lower rate (25 percent). Out of the 62 total correlations that are left over when pairings where correlations are not possible are removed because of partial distributions, only 8 pairings (13 percent) show near-zero or negative correlations consistently across all three provinces.

The variability between provinces and the presence of high rates of near-zero and negative correlations among other provinces shows that the higher incidence of near-zero correlation in the focal provinces when aggregated is not consistent, and not indicative of fabrication.

d. Restricted Ranges

Paper Summary:

The fourth point Koczela and Spagat provide as evidence of potential fabrication is the restricted ranges on interval scale variables among the focal group as compared to the rest of the supervisors. They suggest that fabricators try to avoid outliers, and therefore are less likely to input data that is on the very low or very high end of the scale, resulting in a smaller range than would be found in non-fabricated data.

The authors tested this possibility again by dichotomizing the data into “focal” and “other” supervisors and charting the minimum and maximum value for responses to 12 questions. They show that “the ranges for the focal supervisors are contained within the ranges for the other supervisors in every case.” While this appears to often be true, the minimum values in eight of the 12 charts appear to be identical or nearly identical for both the focal and other groups; and in eight of the 12 variables, focal and other supervisors have the same range in at least one of the polls, meaning that the frequency with which the focal supervisors have restricted ranges appears to be somewhat less than suggested.

Koczela and Spagat also chart the full distribution of the interview duration variable in the five polls, contrasting focal and other supervisors. These charts show that the distributions for the focal supervisors have truncated tails, especially on the right. This is another way of showing that the focal supervisors’ data has less variance than the other data, with less data toward the extremes of the distribution.

However, both of these demonstrations of restricted range among the focal supervisors could be explained simply by the fact that the focal supervisor group includes half the amount of interviews as the other data. The likelihood of having data on the extreme ends of a distribution (i.e., outliers) increases with more interviews. Therefore, the simple fact that the other group was larger in size may be masking the fact that many individual supervisors within that group have ranges that are similar to the ranges found in the focal supervisors’ data.

ABC data:

To test this possibility, we reviewed eight continuous variables in the ABC #1087 dataset: interview start time, interview end time, duration of interview, age, years of education, income, the number of people in the household and the number of people present at the interview. First we checked to see whether the patterns detected in the Koczela and Spagat papers were replicated in our dataset by comparing the range of the focal supervisors (aggregated together) on each variable to the range of the other supervisors (aggregated together).

We then looked at the range for each of these variables among each individual supervisor. If Koczela and Spagat's accusations of potential fabrication were true, we should find that the ranges for these variables among the focal supervisors are smaller than the ranges among the non-focal supervisors at the individual level as well as the aggregate level.

Interview Start Time

For every interview conducted, interviewers were asked to fill in the start time of that interview (e.g., 8:15 a.m.). The easiest way to determine whether focal supervisors show a restriction of range on this variable is to subtract their latest start time from their earliest start time. Replicating Koczela and Spagat, we find that the range for the focal supervisors, collapsed into one group, is 512 minutes, whereas the range for the other supervisors is much larger, 659 minutes.

However, repeating this same analysis separately for each individual supervisor in the focal and other groups reveals a different pattern. The table below shows the range (in minutes) for each supervisor, in order from smallest to largest. Focal supervisors are indicated with an asterisk.

Supervisor number	Start range in minutes
67	219
77	249
80	260
36*	265
29	307
91	339
1	484
93*	488
44*	490
47*	504
94*	504
43*	509
63	519
2	530
6	535
53	542
84	545
58	557
56	558
25	560
32	560

Review of ABC News Iraq data – August 2011

16	565
46	575
88	590
23	621
50	658

As can be seen, contrary to Koczela and Spagat's contention, the focal supervisors did not have the most restricted (i.e., smallest) range of all the supervisors. Indeed six of the seven smallest ranges are for non-focal supervisors, and five of the six focal supervisors had a greater than average range ($M = 482$ minutes).

While more of the other supervisors had larger ranges than the focal supervisors did, there are possible explanations for this that do not suggest data fabrication, e.g. more compressed work hours in densely populated or more dangerous areas. The focal supervisors' interviewers appear to have worked a 9 to 6 schedule; supervisors who worked in more rural areas tended to get an earlier start and worked longer days.

Interview End Time

The same analysis and nearly the same results were obtained when we looked at the ranges for interview end time (i.e., the latest time the interview ended an interview – the earliest they ended an interview, in minutes). The earliest end time recorded was 8:25 a.m., for interviewer 2; the latest end time was 7:32 p.m., for interviewer 50.

In the aggregate, focal supervisors again showed a far smaller range than the other supervisors, 511 vs. 667. However, at the individual level, the focal supervisors did not have the shortest ranges; instead four of the six focal supervisors had above average ranges on this variable ($M = 482.6$). Not surprisingly, the pattern is nearly identical to the interview start time data.

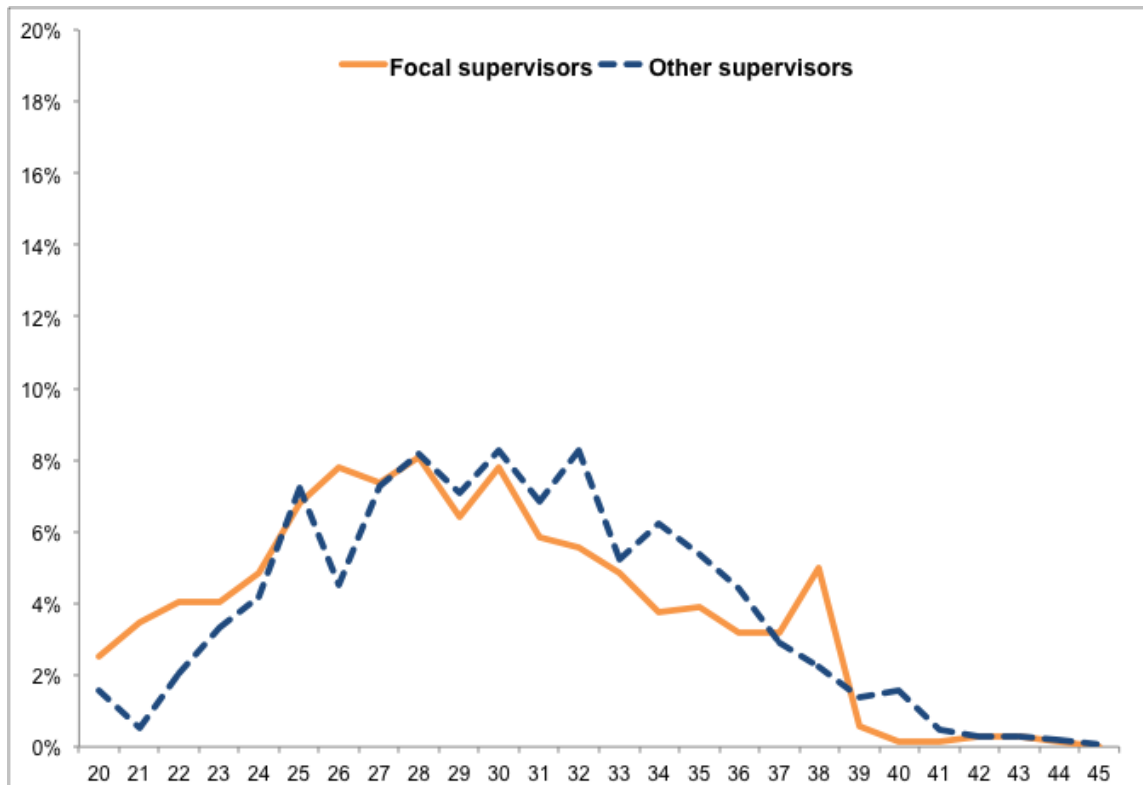
Supervisor number	End range in minutes
67	224
77	245
80	265
36*	276
29	307
91	343
44*	476
1	482
93*	489
94*	495
43*	508
47*	511
63	528
6	528
2	532
53	539
84	545
56	549

Review of ABC News Iraq data – August 2011

25	564
16	565
32	566
58	571
46	576
88	590
23	620
50	654

Duration of Interview

In contrast to interview start and end time, we do not replicate the pattern Koczela and Spagat found for interview duration in the ABC #1087 poll, even at the aggregate level. In the focal supervisors group, the length of interviews ranged from 20 to 44 minutes (a 24-minute span), and in the other supervisor group they ranged from 20 to 45 minutes (a 25-minute span). In addition, as seen below, the full distribution of interview durations is nearly identical for the focal supervisors and other supervisors (contrast with Koczela and Spagat, Figure 4).



As is shown in the table below, at the individual level the focal supervisors did not have anywhere close to the shortest ranges in interview length.

Supervisor number	Range in minutes
53	13

Review of ABC News Iraq data – August 2011

84	14
23	15
67	15
91	15
25	16
56	16
80	16
2	17
16	17
32	18
36*	18
94*	18
1	19
6	19
29	19
93*	19
88	20
47*	21
77	21
46	22
58	22
44*	23
43*	24
63	24
50	25

Age of respondent

In the aggregate, the focal supervisors show a smaller range in respondent age than the other supervisors, replicating Koczela and Spagat. The ages of respondents in the focal supervisor data range from 18 to 69, a span of 51 years, whereas in the other supervisors data, the ages range from 18 to 83, a span of 65 years.

If the focal supervisors were “lacking the ability or imagination to predict outliers or full scale ranges” as Koczela and Spagat suggest, we would expect them also to have the shortest ranges at the individual supervisor level. But as the table below shows, the focal supervisors do not have the shortest age ranges. In fact they tend to have wider age ranges than most other supervisors.

The only reason the other supervisors in the aggregate had such a wide age range is because of one supervisor’s data (Supervisor 46), which includes a respondent aged 83. The rest of the other supervisors had the same or shorter ranges than the focal supervisors.

Supervisor number	Range in years
67	35
53	38
16	39
6	40
2	41

Review of ABC News Iraq data – August 2011

29	41
56	41
91	41
80	42
25	43
36*	43
84	43
88	43
50	45
23	46
43*	46
44*	47
93*	47
32	48
47*	49
63	50
1	51
58	51
77	51
94*	51
46	65

Years of Education

A similar story emerges for years of education. In the aggregate, the focal supervisors' data ranged from 0 to 16 years of education, a span of 16 years, whereas the other supervisors data ranged from 0 to 22 years of education, a span of 22 years.

Disaggregated, however, we show that the other supervisors' wider range is due to just one supervisor's data (Supervisor 53). Most of the other supervisors have education ranges of 16, 17 or 18, about the same as the focal supervisors' data. Indeed, in the entire data set there is just one person who had 22 years of education and three people who had 19 years of education. Ninety-seven percent of the population (unweighted) had 16 years or less of formal education.

Supervisor number	Range in years
63	9
1	16
25	16
29	16
36*	16
43*	16
44*	16
47*	16
80	16
88	16
93*	16
94*	16
58	17
67	17

Review of ABC News Iraq data – August 2011

2	18
6	18
16	18
23	18
32	18
50	18
56	18
77	18
84	18
91	18
46	19
53	22

Household Income

While Koczela and Spagat did not look at household income in their restricted range analysis, it should, theoretically, demonstrate the same pattern if the focal supervisors fabricated their data. Indeed in the aggregate, the other supervisors' data ranges from a minimum income of 100,000 dinars to a maximum of 2,500,000 dinars, a range of 2,400,000, whereas the focal supervisors data ranges from a minimum income of 125,000 to a maximum of 1,100,000, a range of just 975,000. This means there is a 1,425,000 gap between the two ranges.

However, again, a completely different pattern emerges in the individual data. The individual focal supervisors show household income ranges toward the wider end of the spectrum. Only one non-focal supervisor had data with a wider range than the focal supervisors.

Supervisor number	Range in dinars
1	75,000
58	170,000
56	220,000
67	440,000
91	450,000
6	470,000
50	500,000
2	550,000
16	550,000
32	550,000
53	550,000
84	550,000
29	600,000
25	650,000
23	800,000
77	800,000
43*	850,000
93*	850,000
36*	875,000
44*	875,000
94*	900,000

Review of ABC News Iraq data – August 2011

47*	950,000
46	2,350,000

Number of People in Household

Contrary to Koczela and Spagat, we find only a very small discrepancy in range between the focal and other supervisors for number of people in the household. For focal supervisors, between 4 and 17 people lived in the same household, a span of 13 people, whereas for other supervisors, the number of people in the household ranged from 1 to 16, a span of 15 people.

Disaggregated, the evidence of restricted range becomes even thinner. The focal supervisors' data showed a wider range on this variable than many of the other supervisors, as can be seen in the table below. The 10 lowest ranges were among non-focal supervisors.

Supervisor number	Range in people
29	5
16	9
23	9
32	9
53	9
56	9
6	10
46	10
50	10
67	10
36*	11
63	11
91	11
94*	11
1	12
2	12
43*	12
44*	12
47*	12
84	12
88	12
93*	12
25	13
77	13
80	14

Number of People Present

We do not replicate the pattern Koczela and Spagat find for number of people present at the interview in our ABC 1087 poll, even at the aggregate level. Both the focal and other supervisor groups' data show between three and nine people were present at the interview, a span of six people. (It should be noted that the charts in the Koczela and Spagat paper only document

evidence of restricted range on the number of people present at interview in one poll, PIPA Jan 2006).

Disaggregated, too, there is no evidence that the supervisors show a restriction of range on this variable. In fact, all of the focal supervisors have data that span the full six-person range, whereas many of the other supervisors do not have a full range.

Supervisor number	Range in people
29	1
67	1
23	2
77	2
91	2
2	3
16	3
50	3
53	3
56	3
80	3
84	3
88	3
6	4
32	4
63	4
25	5
46	5
58	5
1	6
36*	6
43*	6
44*	6
47*	6
93*	6
94*	6

In sum, Koczela and Spagat contend that focal supervisors had restricted ranges on their continuous variable data because the data were fabricated, and the fabricators lack “the ability or imagination to predict outliers or full scale ranges.” However, the difference in ranges only holds up when the data are grouped into a smaller focal group vs. a larger other group – and not always then. If the focal supervisors had fabricated their data, we should see a restriction of range at the individual supervisor level as well as in the aggregate. Instead, looking at the ranges among all of the supervisors individually reveals that the focal supervisors never have the most restricted range on an interval variable.

Koczela and Spagat write in a footnote that, “since the other supervisors handled more interviews than the focal ones, we might expect some tendency for narrower ranges for the latter compared to the former.” But they then dismiss this idea as insufficient to explain “extremely different ranges.” However, disaggregating the data into individual supervisors shows that the

difference in sample size does in fact appear to explain all of the difference in ranges between the focal and other supervisor groups.

e. Implausible Household Viewership Patterns

Paper Summary:

Finally, Koczela and Spagat describe BBG data that reflects implausible relationships across households in television viewing patterns. Specifically they show that respondents in the focal supervisors' data frequently switch their televisions on for half-hour slots before shutting them off, which is unusual but not unheard of. However, they tend to flip on and off their TVs in a diagonal pattern, such that respondent A watches from 8:00 to 8:30, respondent B watches from 8:30 to 9:00, respondent C watches from 9:00 to 9:30, etc. This is not what one would expect based on random selection of households.

However this pattern also occurs to some extent in the other supervisors' data. The authors offer no summary statistics about how often this occurred, and no statistical tests of whether the two distributions differ. We have no such questions in ABC data to analyze.

Conclusion

We take allegations of data fabrication extremely seriously, hence the extent of the evaluation reported here. We have conducted detailed analyses of partial distributions, DK/Refs, correlations and restricted ranges in ABC Iraq data. We find no evidence of fabrication.