20th International Congress on Modelling and Simulation, Adelaide, Australia, 1–6 December 2013
www.mssanz.org.au/modsim2013

On issues concerning the assessment of information contained in aggregate data using the F-statistic

S.A. Cheema[a], E.J. Beh[a] and I.L. Hudson [a]

[a] *School of Mathematical and Physical Sciences, University of Newcastle, Australia*
Email: salman.cheema@uon.edu.au

Abstract: The analysis of aggregate data has been gaining momentum in the statistics and allied disciplines, (including public policy, political science and epidemiology) for more than 20 years. As a result, the issue has received an increasing amount of attention by categorical data analysts. Performing aggregate data analysis is quickly becoming unavoidable in many situations, especially when individual level data is unavailable. For example, the U.S. Justice Department uses aggregate data to formulate the public policies against racial discrimination, political scientists are always interested in exploring the political or ideological preferences of different demographic groups while social scientists use aggregate data to study the relationship between crime and unemployment. The availability of aggregate data has increased due to strict confidentiality restrictions imposed upon by government and corporate organisations who are reluctant to release individual level information. There is a wealth of contributions on this issue that is available in the ecological inference (EI) literature which considers the association structure between categorical variables (at the individual level) given only the aggregate information. The main difficulty in EI arises due to the loss of information during the process of aggregation and results in aggregation bias. It is also a matter of concern for aggregate data analysts that the interpretation of the parameters from EI models might be entirely different to analogous parameters for the study of individual level data. An alternative strategy to EI is to consider the recently proposed *Aggregate Association Index* (AAI) that allows the analyst to quantify the overall extent of association between two dichotomous variables given only the aggregate, or marginal, information of a 2x2 table. Unlike EI, the AAI does not estimate, or model, the conditional proportions but focuses instead on gauging the extent of association between the variables. The AAI can also be further partition into positive and negative association terms thus enabling the analysts to understand the more likely direction of the association given only the aggregate data. However, the major issue with the performance of AAI is the impact the sample size has on its magnitude. In this paper we investigate the informativeness of the aggregate data for inferring an association exists between the variables of a 2x2 table. This article introduces development of an F-test to determine the statistical significance of the information contained in the aggregate data for inferring a statistically significant association between the variables. Unlike Pearson's chi-squared statistic, the F-statistic is robust to any change in the sample size and depends only on the aggregate information in the contingency table. Thus this statistic provides an opportunity to understand the structure of a 2x2 table without being influenced by sample size. The applicability of this test is demonstrated by using the Selikoff's (1981) asbestosis data which was collected from 1117 insulation workers of New York City in 1963 to explore the links between asbestosis and occupational exposure to asbestos fibres. Such work was the key to establishing the link between asbestosis and mesothelioma. As a result of findings of this nature, many international government organisations have now banned the production, and importation, of goods that contain asbestosis fibres.

Keywords: *Aggregate data, Aggregate Association index, Selikoff's data*

1. INTRODUCTION

The analysis of aggregate data has received a considerable amount of attention from researchers across a broad range of disciplines in the past 20 years. In particular, ecology, statistics and political science have made significant contributions to this area of research. The use of aggregated data is almost unavoidable in many areas of research due in part to the imposition of strict privacy policies by government and commercial.

Discussions concerning the utility of aggregate data when only marginal level data is available has deep statistical roots. Fisher (1935) questioned the usefulness of aggregate data and he was convinced that it is of limited use. Placket (1977) and Berkson (1978) considered the same issue and concluded that aggregate data can be utilized in order to infer about joint frequencies. Yates (1984, pp. 447) agreed with Fisher's conclusion, however, he argued that, for "extreme" marginal frequencies, the estimation of the cell values was possible. Haber (1989) demonstrated that the maximum likelihood estimate of the joint cell values of a 2x2 contingency table do not exist unless one of the cells is zero. Aitkin and Hinde (1984), Bernard (1984) and Beh, Steel and Booth (2002) also made valuable contributions to this discussion.

The most commonly used set of techniques to analyse aggregate data are those belonging to ecological inference (EI) - the growth of which has been strong in the development and application of these techniques. These EI techniques focus on the estimation and modelling of the cell values (or some simple function of them) for multiple or stratified 2x2 contingency tables. Goodman (1953, 1959), Freedman et.al. (1991), King (1997), King, Rosen and Tanner (1999), Steel, Beh and Chambers (2004), Wakefield (2004) and more recently Wakefield, Haneuse, Dobra and Teeple (2011) have made considerable contributions to the growing literature on this topic. Hudson, Moore, Beh and Steel (2010) have provided an extensive discussion and comparison of many of the more popular EI techniques. However, the assumptions that are imposed for these EI techniques are unrealistic, untestable or hard to meet in real life. Therefore, recent contributions on the topic have shifted from modeling the cell frequencies given only the aggregate data (which underlies all EI techniques) to the analysis of the association structure between dichotomous variables. In doing so, Beh (2008, 2010) proposed a new index called the *Aggregated Association Index* (AAI) which quantifies the extent of association that may exist between two dichotomous variables at the α level of significance, given only the aggregate data from a single 2x2 contingency table. However, the magnitude of the AAI is under the influence of the sample size of the contingency table. Beh et. al. (2013) have proposed some adjustments to reduce the effect of sample size on AAI.

In this article we propose an F-statistic that can be used to formally test the statistical significance of the extent of information about the association structure when only the marginal information of a single 2x2 contingency table is available. Our statistic is robust to any change in sample size and its applicability is demonstrated by using Selikoff's (1981) asbestosis data. This paper is further divided into three sections. Section 2, defines the notation used and Beh's (2008, 2010) AAI. The F-statistic is introduced in Section 3 and its application is demonstrated in Section 4. Some final remarks are made in Section 5.

2. THE AGGREGATE ASSOCIATION INDEX

2.1. Notation

Consider a random sample of n individuals, or units, that is cross-classified according to two dichotomous variables to form a 2x2 contingency table. Denote n_{ij}as the joint frequency of the $(ij)th$ cell of this table and $p_{ij} = n_{ij}/n$ the corresponding cell proportion. The marginal cell frequencies for the $i'th$ row and $j'th$ column are $n_{i.} = \sum_{i=1}^{2} n_{ij}$and $n_{.j} = \sum_{j=1}^{2} n_{ij}$, respectively, and their marginal cell proportions are denoted as $p_{i.} = n_{i.}/n$ and $p_{.j} = n_{.j}/n$. The general structure of a 2x2 contingency table is presented in Table 1.

Table 1: A general 2x2 contingency table

	Column 1	**Column 2**	**Total**
Row 1	n_{11}	n_{12}	$n_{1.}$
Row 2	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	n

The expected value of n_{ij} under some criteria (including, but not limited to, independence) is denoted by e_{ij}. For example, under the hypothesis of independence between the two dichotomous variables, $e_{ij} = n_{i.}n_{.j}/n$, while the mean cell frequency is $a_{ij} = n/4 = \bar{n}$. Define $P_1 = n_{11}/n_{1.}$ as the conditional probability of the classification of an individual/unit into "Column 1" given that it has been classified into "Row 1".

When the individual level data is not available in Table 1, the n_{11} is bounded as

$$\mathrm{A}_1 = \max(0, n_{.1} - n_{2.}) \leq n_{11} \leq \min(n_{.1}, n_{1.}) = \mathrm{B}_1. \tag{1}$$

The bound (1) has often been used in the EI literature; see, for example, Duncan and Davis (1953), King (1997), Steel, Beh and Chambers (2004) and Wakefield (2004). This bound can be alternatively expressed in terms of P_1 as

$$L_1 = \max\left(0, \frac{n_{.1} - n_{2.}}{n_{1.}}\right) \leq P_1 \leq \min\left(\frac{n_{.1}}{n_{1.}}, 1\right) = U_1. \tag{2}$$

Beh (2010) showed that bound (2) is narrowed to

$$L_\alpha = max\left(0, p_{1.} - p_{2.}\sqrt{\frac{{\chi_\alpha}^2}{n}\left(\frac{p_{.1}p_{.2}}{p_{1.}p_{2.}}\right)}\right) < P_1 < min\left(1, p_{1.} + p_{2.}\sqrt{\frac{{\chi_\alpha}^2}{n}\left(\frac{p_{.1}p_{.2}}{p_{1.}p_{2.}}\right)}\right) = U_\alpha. \tag{3}$$

when a test the of association is made at the α level of significance. It is important to note that (3) depends only on the marginal information and sample size of the data.

2.2. The AAI

Instead of estimating the cell values by modeling the aggregate data, Beh (2008, 2010) proposed the AAI to quantify the overall extent of association between two dichotomous variables at the α level of significance, given only the aggregate data. The AAI uses the transformation of Pearson's traditional chi-squared statistic in terms of P_1 as

$$X^2(P_1|p_{1.}, p_{.1}) = n\left(\frac{P_1 - p_{.1}}{p_{2.}}\right)^2\left(\frac{p_{1.}p_{2.}}{p_{.1}p_{.2}}\right). \tag{4}$$

Note here that Yates' continuity correction is not incorporated into the statistic; refer to Beh (2010) for a discussion of its exclusion from the AAI calculation. By using (4), the AAI is calculated by

$$A_\alpha = 100\left(1 - \frac{[(L_\alpha - L_1) + (U_1 - U_\alpha)]{\chi_\alpha}^2 + \int_{L_\alpha}^{U_\alpha} X^2(P_1|p_{1.}, p_{.1})dP_1}{\int_{L_1}^{U_1} X^2(P_1|p_{1.}, p_{.1})dP_1}\right). \tag{5}$$

The AAI considers the uniform distribution of every value of P_1across the curve – referred to as the AAI curve – defined by (4). This curve is depicted in Figure 1.The AAI ranges between 0 and 100 where a near zero AAI indicates that, given the aggregate data, there is no evidence of a statistically significant association existing between the dichotomous variables. However values of the AAI close to 100 indicate that there exists considerable evidence to conclude that such a statistically significant association exists. A graphical depiction of the AAI can be seen by viewing the proportion of the total area under the curve – defined by (4) – that is shaded in Figure 1.

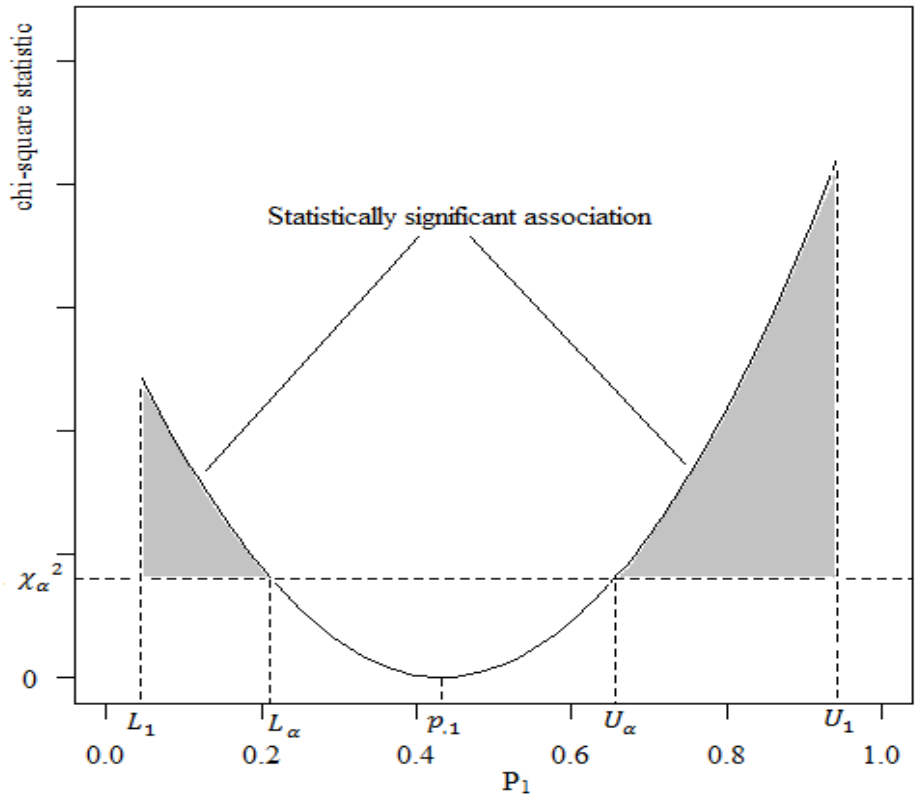

Figure 1: A visual display of AAI; its magnitude is the proportion of the area under the AAI curve that is shaded

2.3. The Least Informative Situation (LIS)

The major issue with the AAI is that its magnitude is under the influence of the sample size. For example, as the sample size increases the value of AAI also increases. Beh (2010) recognized that, when the sample size is equally distributed across the row and column marginals such that $n_{i.} = n_{.j} = \frac{n}{2}$, then there exists the least information available in the marginal data to infer anything about the statistical significance of the association between the variables. We refer to this situation as the least informative situation, or LIS, and it can be used in the development of procedures to formally test the statistical significance of the information in the aggregate data. Interestingly, in the LIS, the expected cell values under independence are equal to the average cell value so that $e_{ij} = a_{ij}$ for all i and j. In the LIS equation (2) can be simplified to

$$L_1 = 0 \leq P_1 \leq 1 = U_1. \tag{6}$$

and (4) can be written as

$$X_L{}^2(P_1) = X^2(P_1 | p_{1.} = \tfrac{1}{2}, p_{.1} = \tfrac{1}{2}) = n(2P_1 - 1)^2. \tag{7}$$

where $X_L{}^2(P_1)$ refers the AAI curve defined by (4) when the LIS is observed.

3. METHODOLOGY

3.1. Degrees of Freedom

As it is well understood, when considering the test of independence between two dichotomous variables, Pearson's chi-squared statistic is a chi-squared random variable with one degree of freedom. As a result, for the development of AAI, Beh (2008, 2010) also considered one degree of freedom. However, when only aggregate data is available, there are five pieces of information known to the analyst; these are the four marginal proportions and the total sample size. Therefore, we need only three values to complete the

structure at marginal level; any of the two row marginal values, any of the two column marginal values and the sample size. We therefore consider that every point along the AAI curve graphically depicted using (4) follows a chi-square distribution with *three* degrees of freedom instead of Beh's (2008, 2010) one degree of freedom.

Since all that is required for the LIS is the sample size under this situation we consider the chi-squared distribution with one degree of freedom for every possible value of P_1 across the AAI curve depicted using (7).

3.2. Test Statistic

Here we introduce a formal procedure for testing the statistical significance of the extent of information given only the aggregate data of a 2x2 contingency table. Suppose we consider the following test:

H_O: The aggregate data is not informative about the association structure of a 2x2 table.

H_A: The aggregate data is informative about the association structure of a 2x2 table.

When undertaking such a test, we should consider that the test statistic reflects the maximum deviation of the marginal structure of a 2x2 contingency table from the LIS. In doing so, the overall mean value of Pearson's chi-squared statistic across the possible range of P_1 values is obtained by integrating (4) over the range (2) and then dividing this by (2) yielding

$$\frac{\int_{L_1}^{U_1} X^2(P_1|p_{1.},p_{.1})d_{P_1}}{U_1-L_1} = \frac{nk^2[(U_1-p_{.1})^3-(L_1-p_{.1})^3]}{3(U_1-L_1)}. \tag{8}$$

The sampling distribution of statistic in equation (8) follows a chi-square distribution with three degrees of freedom, where, $k^2 = p_{1.}/(p_{.1}p_{.2}p_{2.})$. For the LIS, equation (8) may be simplified to

$$\int_0^1 {X_L}^2(P_1)d_{P_1} = \frac{2n}{3}. \tag{9}$$

since, in this case, $L_1 = 0$ and $U_1 = 1$. The statistic (9) follows a chi-square distribution with one degree of freedom. It is worth noting that both statistics, (8) and (9), are independent to each other as they are based upon different chi-square distributions and degrees of freedom. Thus, by considering the ratio of (8) and (9) and dividing them by their respective degrees of freedom, the general expression of the F-test statistic is

$$F = \frac{6(U_1-L_1)}{k^2[(U_1-p_{.1})^3-(L_1-p_{.1})^3]}. \tag{10}$$

Equation (10) follows the F-distribution with numerator degrees of freedom 1 and denominator degrees of freedom 3. It is apparent by considering the F-statistic (10) that it does not depend on the sample size and its magnitude depends only on the available aggregate data.

4. APPLICATION

The applicability of F-statistic given by equation (10) for testing the information contained in the marginal proportions for inferring a statistically significant association between the dichotomous variables is established here by using Selikoff's (1981) asbestosis data set. In 1963, a study was conducted that involved 1117 insulation workers in New York. This study, and its findings published by Selikoff (1981), established the link between long-term occupational exposure to asbestos fibers and the severity of asbestosis the workers were diagnosed with. The impact of this study data established for the first time the links between asbestos exposure and lung disease. As a result, many governments have introduced laws that ban the manufacturing, and importation, of products containing asbestos fibres. This data, summarized in Table 2, has also been a topic of statistical discussion by Beh and Smith (2011) and Tran, Beh and Smith (2012); the latter studied the data in terms of the AAI. Table 2 also summarises the expected cell frequencies under independence between the variables (in parentheses on the left hand side) and under the LIS (in parentheses on the right hand side).

Table 2: Diagnosis and exposure to asbestosis

Selikoff's asbestosis data				Least informative situation (LIS)			
	Asbestosis				Asbestosis		
Onset of exposure	**No**	**Yes**	**Total**	**Onset of exposure**	**No**	**Yes**	**Total**
0-19 years	522*(373.21)	203(351.79)	725	0-19 years	(279.25)	(279.25)	558.5
20+ years	53(201.79)	339(190.21)	392	20+ years	(279.25)	(279.25)	558.5
Total	575	542	1117	**Total**	558.5	558.5	1117

*Observed cell frequencies. Expected cell frequencies under the independence are in the brackets

Pearson's chi-squared test of independence shows that there is a statistically significant association between the time spent exposed to asbestos fibres and whether a work contracts asbestosis (p-value < 0.0001). For Table 2, $(p_{1.}, p_{2.}) = (0.65, 0.35)$ and $(p_{.1}, p_{.2}) = (0.51, 0.49)$ and thus by using (2), the bounds of P_1 for the AAI curve is $0.25 \leq P_1 \leq 0.79$. The AAI for Table 2 is 99.81, which is very high. The magnitude of the index therefore indicates that it is very likely, based only on the aggregate data, that a statistically significant association exists between the dichotomous variables if a test of association is made at the 5% level of significance; thus confirming the findings of the chi-squared test of independence (where the individual level information is known). However, the magnitude of the AAI may be because the marginal information suggests that a statistical significance between the variables is very likely, or perhaps it's because of the moderately high sample size of n = 1117. To explores this, the F-statistic (10) is considered.

Equations (4) and (7) are calculated as

$$X^2\left(P_1 | p_{1.} = \frac{725}{1117}, p_{.1} = \frac{575}{1117}\right) = \frac{878617169}{106232}\left(P_1 - \frac{725}{1117}\right)^2$$

and

$$X_L{}^2(P_1) = 4468\left(P_1 - \tfrac{1}{2}\right)^2,$$

respectively. By using ranges (2) and (6), the value of expressions (8) and (9), are 20.49 and 744.67 respectively. Therefore, the F-statistic (10) is 106.02 and, for a F-distributed random variable with 1 and 3 degrees of freedom, has a p-value of 0.002.This small p-value provides enough evidences against the null hypothesis, leading to the conclusion that the aggregate data is in fact useful for assessing the statistical significance of the association structure between time of exposure and presence of asbestosis.

5. DISCUSSION

The recently developed AAI quantifies how likely a statistically significant association will exist between two dichotomous variables of a 2x2 contingency table given only the aggregate data of the table. Here we have presented a new approach to test whether the magnitude of the AAI is due to the informativeness of the aggregate data in making such a conclusion, or whether its perhaps due to the sample size considered. The key to this test is the F-test statistic defined by equation (10). The advantage of considering this statistic is that conclusions from such tests are independent of the sample size thus helping to establish the statistical significance of the association structure between dichotomous given only the aggregate data. The applicability of the statistic is demonstrated by using the Selikoff's (1981) asbestosis data.

Future research into this aspect of aggregate data analysis can be made by developing an extension of the AAI and the F-statistic for contingency tables of size larger than 2x2. Adapting these measures for multiple, or stratified, 2x2 tables can also be considered. For larger dimensional contingency tables, formed by cross-classifying ordered categorical variables, an interesting extension is the development of the AAI and its F statistic taking into consideration the structure of these variables. On the basis of the LIS, a further issue that can be investigated is the development of a test of homogeneity of the information among strata for stratified 2x2 tables.

REFERENCES

Aitkin, M. and Hind, J. P. (1984). Comments to Tests of significance for 2x2 contingency tables, *Journal of Royal Statistical Society, Series A*, 47, 453 – 454.

Barnard, G. A. (1984). Comments to Tests of significance for 2x2 contingency tables, *Journal of Royal Statistical Society, Series A*, 47, 449 – 450.

Beh, E. J. (2008). Correspondence analysis of aggregate data: The 2x2 table, *Journal of Statistical Planning and Inference*, 138, 2941 – 2952.

Beh, E. J. (2010). The aggregate association index, *Computational Statistics and Data Analysis*, 54, 1570 – 1580.

Beh, E. J. and Smith, D. R. (2011). Real world occupational epidemiology, Part 1: Odds ratios, relative risk and asbestosis, *Archives of Environmental and Occupational Health*, 66, 119 – 123.

Beh, E. J., Steel, D. G. and Booth, J. G. (2002). What useful information is in the marginal frequencies of a 2x2 table?, University of Wollongong Preprint 4/02.

Beh, E.J., Cheema, S.A., Tran, D. and Hudson, I.L. (2013). Adjusting the aggregate association index for large samples, *Italian Statistical Society: Statistical Conference*, University of Brescia.

Berkson, J. (1978). In dispraise of the exact test: Do the marginal totals of the 2x2 table contain relevant information respecting the table proportion, *Journal of Statistical Planning and Inference*, 2, 27 – 42.

Chambers, R.L., Steel, D.G., (2001). Simple methods for ecological inference in 2 × 2 tables. *Journal of the Royal Statistical Society, Series A*, 164, 175 – 192.

Duncan, O. D. and Davis, B. (1953). An alternative to ecological correlation, *American Sociological Review*, 18, 665 – 666.

Fisher, R. A. (1935). The logic of inductive inference (with discussion), *Journal of Royal Statistical Association, Series A*, 98, 39 – 82.

Freedman, D. A., Stephen, P. K., Jerome, S., Charles, A. S. and Charles, G.E. (1991). Ecological regression and voting rights. *Evaluation review* 15, 673-711.

Goodman, L. (1953). Ecological regressions and the behaviour of individuals, *American Sociological Review*, 18, 663 – 666.

Goodman, L. (1959). Some alternatives to ecological correlation, *The American Journal of Sociology*, 64, 610 – 625.

Haber, M. (1989). Do the marginal total of a 2x2 contingency table contain information regarding the table proportion, *Communication in Statistics: Theory and Methods*, 18, 147 – 156.

Hudson, I.L. Moore, L. Beh, E.J. Steel, D.G. (2010). Ecological inference techniques: an empirical evaluation using data describing gender and voter turnout at New Zealand elections 1893-1919. *Journal of the Royal Statistical Society*: Series A, 173, 185 – 213.

King, G. (1997). *A Solution to Ecological Inference Problem*. Princeton University Press: Princeton, U.S.A

King, G., Rosen, O. and Tanner, M. A. (1999). Binomial-beta hierarchical models for ecological inference, *Sociological Methods & Research*, 28, 61 – 90.

Plackett, R. L. (1977). The marginal totals of a 2x2 table, *Biometrika*, 64, 37 – 42.

Selikoff, I. J. (1981). Household risk with inorganic fibres, *Bulletin of the New York Academy of Medicine*, 57, 947 – 961.

Steel, D. G., Beh, E. J. and Chambers, R. L. (2004). The information in aggregate data. In *Ecological Inference: New Methodological* Strategies, (ed King, G., Rosen, O. and Tanner, M. A.), 51 – 68.

Tran, D.Beh, E. J. Smith, D. R. (2012). Real-World Occupational Epidemiology, Part 3: An Aggregate Data Analysis of Selikoff's "20-Year Rule", *Archives of Environmental and Occupational Health*, 67, 243 – 248.

Wakefield, J. (2004), Ecological inference for 2x2 tables, *Journal of the Royal Statistical Society A*, 167, 385 – 445.

Wakefield, J., Haneuse, S., Dobra, A. and Teeple, E. (2011), Bayes computation for ecological inference, *Statistics in Medicine*, 30, 1281 – 1396.

Yates, F. (1984). Tests of significance for 2x2 contingency tables (with discussion), *Journal of Royal Statistical Society, Series A*, 147, 426 – 463.