METHODOLOGY NOTE ON THE CALCULATION OF STATISTICAL WEIGHTS

This note describes the procedure for calculating stratum-specific weights or multipliers for the South African Business Innovation Survey (BIS). This procedure was devised by Prof. Tim Dunne and subsequently modified by Mr Stephen Davis. It was adapted for the BIS 2019-2021 by Dr Moses M. Sithole. The notion of weights assumes that each respondent in the database represents a greater number of virtual firms in the entire population, as governed by the principle of random sampling.

The sampling frame for the BIS is characterised by six main sectors, which are numbered according to their Standard Industrial Classification (SIC) codes. These were further subdivided into sub-sectors as follows: Sector 2 (mining and quarrying: six subsectors), Sector 3 (manufacturing: 10 subsectors), Sector 4 (electricity, gas and water supply: two subsectors), Sector 6 (wholesale trade and retail trade: two subsectors), Sector 7 (transport, storage and communications: five subsectors), and Sector 8 (financial intermediation, computer and related activities, R&D, architectural and engineering activities and technical testing and analysis: five subsectors).

Additionally, each of the subsectors is divided into four distinct size classes, which are classes of firms defined by firm size as described in the accompanying Survey Methodology Note. The most magnified view therefore reveals a frame consisting of (6+10+2+2+5+5) x 4 = 30 x 4 = 120 strata by subsector and size. The following explanation holds for an individual stratum.

The aim of the weighting methodology is to provide a reasonable estimate of the number of innovation-active firms and the number of non-innovation active firms in each stratum, so that appropriate weights can be assigned to each entity in the database in accordance with its status as an innovation-active firm or non-innovation active firm. The mathematical symbols (in italics) refer to the values within each of the 120 strata. Explanations are given below on how the underlying population was determined. From the population we can then infer the portion of non-respondents. Estimates of numbers of innovation-active firms and non-innovation active firms are calculated for each stratum, followed by a revision

of the numbers to provide for those strata in which no weight could have been calculated. Final weights are calculated by dividing the multipliers in each stratum by the actual number of respondents in that stratum.

Reduced population

The presence of non-valid companies in the sample served as a basis to reduce the original population size by a proportion, *p*, to *pN* = valid population. An important assumption was that the original population consisted of a number of untraceable or expired companies that are no longer part of the base population. The procedures make **no allowance for growth** in the population size in the form of new entrants. Resultant estimates will on balance tend to be conservative and a "good" guess of baseline levels of innovation activity. Valid population sizes were therefore calculated from the proportion of valid firms as ascertained from the fieldwork.

Non-response

Of the total (valid) sample in each stratum, a proportion, *q*, responded to the survey. So we infer a population of respondents in each stratum of *R = qpN*, and a corresponding population of non-respondents, *NR = (1-q)pN*.

Innovation-active and non-innovation active firms

Of the valid respondents, a proportion, *k*, were found to be innovation-active, so that the number of innovation-active firms among the *R* responders is *I1 = kR = kqpN*. To estimate the innovation-active rate among non-responders, a non-response survey was conducted in the form of a simple random sample (SRS). The SRS revealed a 62.9% innovation-active rate amongst the total *NR*. Therefore, the total number of virtual innovation-active firms among the non-responders is *I2 = NR x 62.9%*. **A vital assumption was the 62.9% rate of innovation activity assumed amongst the estimated number of non-respondents in <u>all</u> strata**. Within each of the 120 strata, the total estimated population of innovation-active firms is *I = I1 + I2*. Implicitly, we therefore have a number of non-innovation active firms, *NI = pN – I*, in each stratum.

Missing weights

The problem of zero weight, where no estimate *k* was calculable since no responses were returned, only arose in two strata in the crude petroleum and natural gas subsector within the mining and quarrying sector. The consequence is that all the valid firms, R, in the corresponding strata were implicitly assumed to be non-innovation active. As similar strata in the main sector and size class did actually contain a portion of innovation-active firms, there was a potential undercounting of innovation-active firms overall (and a subsequent over-counting of non-innovation active firms, and therefore overstating of the true value of *NI*).

Revised weights

An option for revising the weights is to assume the 62.9% innovation-active rate in the non-response survey applies to the valid entities in the "weightless" strata and enter *pN x 62.9%* for these. A more internally consistent method is to inflate the (weighted) strata of the same size class within the sector by the proportion missing or "weightless" for that sector and size class. In this way, some of the homogeneity of sector and size class is maintained. The method is illustrated in an example below (Table 1). Where applicable, we then calculate *I** and *NI**, the inflated number of innovation-active and non-innovation active firms in the target population respectively.

Final weights

For each stratum, two types of weights are then calculated. The first, *W1*, is equal to *I* / n1*, where *n1* is the number of innovation-active firms in the sample. Defined in this way, *W1* will have to be zero for strata where *n1 = 0* and *n2>0*. In the situation where both n1 and n2 equal zero, then the inflation factor kicks in and we inflate the remaining strata with the same size class in the sector. In this manner, any estimates of population totals inferred from the innovation-active firms' data will be conservative, and for certain strata, no totals will emerge. Similar logic applies to the calculation of the second type of weight, *W2*, for non-innovation active firms, which is equal to *NI* / n2*, where *n2* is the number of non-innovation active firms in the sample. When the sample of innovation-active or non-innovation active firms is small and sparsely scattered throughout the 120 strata, we invoke "weight totalling" as shown in Table 1.

Table 1: Illustrative example (Sector 2 and the associated strata)

Strata	pN	n1	n2	I	NI	I + NI	Missing entities	Inflation factor	(I + NI)*	I*	NI*	W1	W2
21 1	72	8	9	42.57655	29.42345	72	0		72	42.57655	29.42345	5.314924	2.602047
21 2	33.3	1		22.30847	10.99153	33.3	0		33.87559	22.69407	11.18152	11.25041	
21 3	18	1	1	10.8013	7.198697	18	0		18	10.8013	7.198697	15.42915	9.203135
21 4	68		1	37.40554	30.59446	68	0		69.16334	38.04547	31.11787		83.1456
22 1	2		1	0.628664	1.371336	2	0		2	0.628664	1.371336		2.602047
22 2	3					0	3		0	0	0		
22 3	5		2	1.885993	3.114007	5	0		5	1.885993	3.114007		9.203135
22 4	16.8		0			0	16.8		0	0	0		
23 1	23	2	4	12.6873	10.3127	23	0		23	12.6873	10.3127	5.314924	2.602047
23 2	2	1		1.628664	0.371336	2	0		2.03457	1.656816	0.377754	11.25041	
23 3	3		1	1.257329	1.742671	3	0		3	1.257329	1.742671		9.203135
23 4	24	1	1	14.31596	9.684039	24	0		24.41059	14.56088	9.849713	194.3573	83.1456
24 1	48	4	8	26.63192	21.36808	48	0		48	26.63192	21.36808	5.314924	2.602047
24 2	24.8	2		17.89316	6.90684	24.8	0		25.22867	18.20244	7.026225	11.25041	
24 3	12.8	1	1	7.635179	5.164821	12.8	0		12.8	7.635179	5.164821	15.42915	9.203135
24 4	91.8	1	1	55.08664	36.71336	91.8	0		93.37051	56.02906	37.34145	194.3573	83.1456
25 1	36	3	11	16.83062	19.16938	36	0		36	16.83062	19.16938	5.314924	2.602047
25 2	78.46154	2	2	45.96091	32.50063	78.46154	0		79.81774	46.75534	33.0624	11.25041	16.01436
25 3	101	2	3	51.94756	49.05244	101	0		101	51.94756	49.05244	15.42915	9.203135
25 4	759		1	443.0738	315.9262	759	0		771.9849	450.6538	321.3311		83.1456
29 1	27	5	2	17.57329	9.42671	27	0		27	17.57329	9.42671	5.314924	2.602047
29 2	35	4	2	22.8013	12.1987	35	0		35.60497	23.19542	12.40955	11.25041	16.01436
29 3	26.4	2		19.04756	7.352443	26.4	0		26.4	19.04756	7.352443	15.42915	
29 4	39.2	1	1	23.38274	15.81726	39.2	0		39.87063	23.78277	16.08786	194.3573	83.1456
Totals													
2 1	**208**	**22**	**35**	**116.9283**	**91.07166**	**208**	**0**	**0.00%**	**208**	**116.9283**	**91.07166**	**5.314924**	**2.602047**
2 2	**176.5615**	**10**	**4**	**110.5925**	**62.96903**	**173.5615**	**3**	**1.7285%**	**176.5615**	**112.5041**	**64.05745**	**11.25041**	**16.01436**
2 3	**166.2**	**6**	**8**	**92.57492**	**73.62508**	**166.2**	**0**	**0.00%**	**166.2**	**92.57492**	**73.62508**	**15.42915**	**9.203135**
2 4	**998.8**	**3**	**5**	**573.2646**	**408.7354**	**982**	**16.8**	**1.7108%**	**998.8**	**583.072**	**415.728**	**194.3573**	**83.1456**
2 Total	1549.562	41	52	893.3604	636.4011	1529.762	19.8		1549.562	905.0794	644.4822	22.07511	12.39389

Looking at Table 1 we note that strata 22 2 and 22 4 have 3 and 16.8 missing entities, respectively. In size class 2, one firm represents 1.7285% of the *I + NI*. We must therefore inflate the non-missing strata (where either *n1 > 0* or *n2 > 0* or both *n1 > 0* and *n2 > 0*) belonging to this size class in this subsector (in this case 21 2, 23 2, 24 2, 25, 2 and 29 2) by 1.7285%. Hence, the increase from 33.3 to 33.87559 evident in stratum 21 2, for example. Similarly, to compensate for the missing 16.8 entities in 22 4, we increase by 1.7108% the 68, 24, 91.8, 759 and 39.2 in 21 4, 23 4, 24 4, 25 4 and 29 4 to 69.16, 24,41, 93.37, 771.98 and 39.87, respectively.

Naturally, we also apply the inflation of the non-missing strata described above to the corresponding innovation-active and non-innovation active firms. Ideally, to calculate the final weights for each stratum, *W1* and *W2*, we should then divide the resulting estimates of target firm population size by the corresponding numbers of sample responses, *n1* and *n2*, respectively.

However, either *W1* or *W2* is incalculable if innovation-active or non-innovation active responses are missing in some strata, i.e., either *n1 = 0* or *n2 = 0*, and using stratum level weights calculated as described above would exclude the estimates of the target firm population size for the strata where this absence occurs from the sector total.

As a result, to each stratum, we assign final weights *W1* and *W2* calculated with the numbers of firms (estimates of strata target population sizes for innovation-active and non-innovation active firms, and *n1* and *n2*) aggregated to size class level. In other words, each non-missing stratum belonging to a particular size class in this sector, say, size class 1, received final weight *W1* and *W2* calculated at size class level for that size class.

For example, innovation-active firms in strata 21 1, 21 2 and 21 3 were assigned weights 5.31, 11.25 and 15.42, which were the size class level weights for size classes 1, 2 and 3, respectively. Similarly, non-innovation active firms in strata 21 1, 21 3 and 21 4 were assigned weights 2.60, 9.20 and 83.15, which are the size class level weights for size classes 1, 3 and 4, respectively.

There were no missing responses for Sector 6: wholesale and retail trade sector, i.e., both *n1* and *n2* were greater than zero for all the strata in this sector. Therefore, the inflation adjustment described above was not applicable for this sector and all the strata level weights were calculable, and hence were calculated and used. Although no inflation adjustment was required in all the other sectors (except for Sector 2, as discussed), either *n1* or *n2* was equal to zero for some strata, which necessitated the aggregation of target firm population estimates and responses to size class level and hence the calculation and use of size class level weights.

If the responses are so sparsely populated that missing innovation-active or non-innovation active responses occur, i.e., either $n1 = 0$ or $n2 = 0$, even after aggregation to size class level, then this would necessitate further aggregation to sector level, and hence the calculation and use of sector level weights, which are not as specific and accurate as size class level or stratum level weights.

During the fieldwork, a great deal of effort went into ensuring that at least one response was realised in all strata. Although this was not achieved in Sector 2: mining and quarrying, the strategy largely worked. The strata in all sectors realised enough responses not to require aggregation of estimates of target firm population sizes and responses to sector level and no sector level weights were used in the survey.