

USING SECONDARY DATA TO DEVELOP SCIENCE REPORT WRITING SKILLS

NIGEL DIXON

XJTU LANGUAGE CENTRE



Xi'an Jiaotong-Liverpool University

西交利物浦大學

EAP103: ENGLISH AND STUDY SKILLS FOR SCIENCE

Programme: Year 2; 2 semesters (10 credits)
2X2-hour seminars/week + self-study

125 students 2018-19, majoring in:

- Biosciences
- Bioinformatics
- Chemistry
- Environmental Science
- Public Health

Assessments:

- Writing Coursework
- Speaking test
- S1 & S2 Exams (Reading, Listening, Essay writing)



COURSEWORK HISTORY: 2016-17

S1 Coursework: Experiment Report (12-1600 words)

Introduction with literature review, method, results, discussion, references

S2 Coursework: Essay (1000-1350 words)

- *S1 Experiment based on US science fair project ideas (www.sciencebuddies.org)*
 - *E.g. How acids affect the rate of corrosion*
 - *How different soil affects the PH of water, etc.*

X Academic departments objected to experiments on safety grounds

X Excessive demand on students' time for non-language tasks

X Students struggled with focus of Experiment Report literature review

X S2: Some students requested focus on reports rather than essays



ASSESSMENT HISTORY: 2017-18 / 2018-19

Semester 1 Coursework

- **Planning task 1: Results & Discussion:** *graphs + description; discussion of results*
- **Planning task 2: Introduction:** *Literature review, Aims/Research Questions*

Semester 2 Coursework

- **Full Report:** *Abstract, Introduction (with Literature review), Method, Results, Discussion, Conclusion, References*
- Students revise, expand and integrate S1 Planning tasks

Objectives

- *Avoid safety / ethical issues*
- *Enhance learning through a focus on writing process and formative feedback (Seviour, 2015)*
- *Students engage with authentic data → Secondary data approach*



RESEARCH BASED ON SECONDARY DATA

Secondary Data: “Data originally collected for a different purpose and reused for another research question” (Hox and Boeije, 2005)

“[T]he use of secondary data is crucial in the education of undergraduate and graduate students (Fienberg 1994; King 2006). It is not feasible for students in a semester-long course to collect and analyze data on a large scale. Using datasets that have been archived and shared allows students to experience science first hand”
Pienta, Alter and Lyle (2010).

In social sciences, student collection of their own research data “is also usually prefaced by an internal, ethical review board. The process can take up valuable class time and place conditions on research activities”

Haaker & Morgan-Brett (2017).



TASK DEVELOPMENT PROCESS

1. Summer 2017: Consultation with departments:
 - Environmental Science: suggested biodiversity e.g. bird populations
 - Biosciences: provided lung cancer patient data set (ICGC, 2017)
2. Research topics and sources of data sets:
3 options offered:
 - 1) Data for 50 cancer patients in Excel spreadsheet (ICGC, 2017)
 - 2) London bird population data (BTO, 2017): data and graphs
 - 3) Personal nutrition 3-day record: student's self-collected data
3. Check availability of background sources (sufficient for literature review?)
4. Design supporting lessons
 - Language needs e.g. describing data and trends
 - Textual analysis of different report sections
 - Research skills: finding and evaluating sources in the literature



2017-18 LUNG CANCER DATA SET

Selected Data for 50 North American lung cancer patients (ICGC, 2017, <https://icgc.org>)

Patient gender	Age at which disease was first diagnosed	Indicates if a person has a history of smoking	Lifetime tobacco exposure: cigarettes per day x no. of years ÷ 20	a DNA sequence resulting in a shorter, unfinished protein product	Disease type
GENDER	AGE	SMOKING_HISTORY	SMOKING_PACK_YEARS	NONSENSE_MUTATIONS	CANCER_TYPE
MALE	67	Current reformed smoker for < or = 15 years	200	275	Non-Small Cell Lung Cancer
MALE	72	Current reformed smoker for > 15 years	40	224	Non-Small Cell Lung Cancer
FEMALE	63	Current smoker	50	353	Non-Small Cell Lung Cancer
MALE	52	Current reformed smoker for > 15 years	25	245	Non-Small Cell Lung Cancer
MALE	73	Current reformed smoker for < or = 15 years	25	383	Non-Small Cell Lung Cancer
MALE	77	Current reformed smoker for > 15 years	30	171	Non-Small Cell Lung Cancer
MALE	83	Current reformed smoker for > 15 years	40	528	Non-Small Cell Lung Cancer
MALE	65	Current reformed smoker for > 15 years	25	241	Non-Small Cell Lung Cancer
MALE	57	Current reformed smoker for > 15 years	40	254	Non-Small Cell Lung Cancer
MALE	50	Lifelong Non-smoker		192	Non-Small Cell Lung Cancer
FEMALE	60	Current reformed smoker for < or = 15 years	15	448	Non-Small Cell Lung Cancer
MALE	67	Current smoker	90	163	Non-Small Cell Lung Cancer
FEMALE	54	Current reformed smoker for > 15 years	35	265	Non-Small Cell Lung Cancer
MALE	77	Current reformed smoker for < or = 15 years	9	62	Non-Small Cell Lung Cancer
MALE	69	Current reformed smoker for < or = 15 years	45	162	Non-Small Cell Lung Cancer

etc. (50 rows)



2017-18 LONDON BIRD POPULATION DATA

Species	Sample	Change 15-16	Change 95-15	Sig	LCL	UCL
Feral Pigeon	73	-7	-24	*	-36	-4
Woodpigeon	82	1	56	*	22	93
Collared Dove	52	-7	-10		-34	26
Magpie	81	-2	36	*	15	65
Jay	40	-14	-14		-34	15
Carrion Crow	82	12	46	*	11	90
Starling	79	-15	-57	*	-68	-46
Blackbird	82	4	-28	*	-38	-15
Song Thrush	51	21	-35	*	-50	-17
Mistle Thrush	34	20	-53	*	-70	-34
Robin	79	7	84	*	52	121
House Sparrow	69	1	-73	*	-80	-66

Columbidae

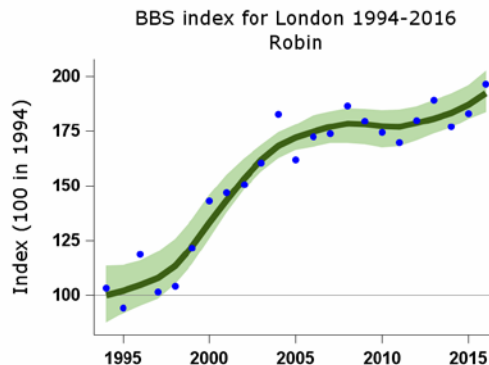
Corvidae

Turdidae

Red-listed
(threatened species)



robin

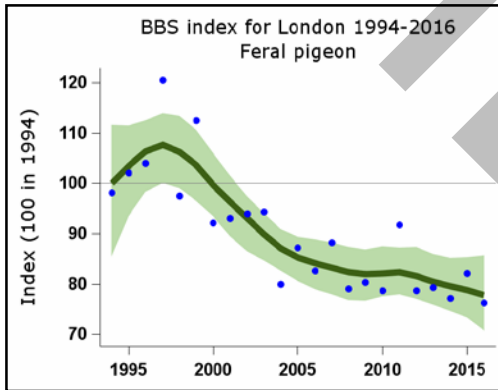


+ 11 more graphs from British Bird Survey website (BTO, 2017)



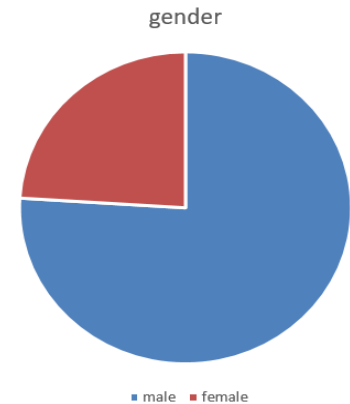
2017-18 THE STUDENT RESPONSE

Based on completed tasks, in class comments & survey responses

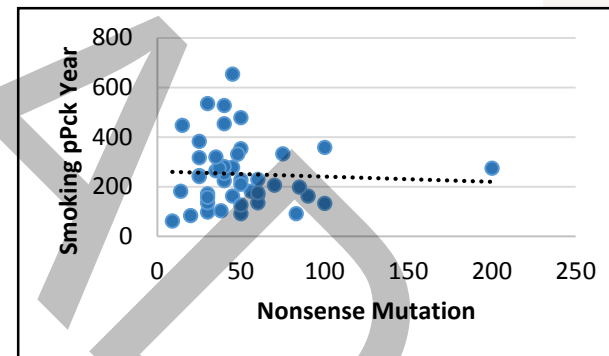


Is this good enough?

What should I do? The graph has already been made.



I don't understand my data.
Is it significant?



What is “smoking pack years”? / “current reformed smoker”? / “upper confidence limit”?

It's confusing to write the Results before the Introduction.



2018-19 GUIDELINES FOR CREATING TASKS

1. Balance authenticity of the input with accessibility:

- A higher number of samples is more meaningful to students
- But simplify or regularize data e.g. remove samples with missing data
- Replace technical vocabulary, or provide a clear glossary
- Avoid data variables which require complex technical knowledge

2. Task options require similar amounts of processing:

- Students create their own graphs for all options
- Raw data provided for all options
- Data include sufficient variables to offer choices, but not be overwhelming

3. Include more in class training in processing raw data

4. Adapt tasks so that research sequence appears more logical

5. Add formative tasks to scaffold students in problematic areas:

- PT 1: graph content and design; confusion over aims
- PT 2: poor referencing format; insufficient academic sources



2018-19 COURSEWORK TASKS

Option 1: **Pancreatic Cancer Patient Data** for 142 Canadian patients; data source: ICGC (2017)

Option 2: **Offshore fisheries data** for seven species, 1970-2014); data source: UN FAO (2018)

Option 3: **Nutritional content of food from two fast food restaurants** (KFC, MacDonald's and/or Subway):

Student choice of data, but two recommended sources: <http://calorielab.com>;
<https://www.verywellfit.com>



2018-19: PANCREATIC CANCER PATIENT DATA

patient sex	Age at diagnosis	tumour stage at diagnosis (TNM system)	Current vital status	Disease status at last follow-up	Relapse type	Relapse-free interval (days)	Patient survival time(days)	cancer history in first degree relative
female	68	IIB	alive	complete remission			1264	yes
male	74	III	deceased	progression			131	unknown
male	68	IIB	deceased	progression			371	yes
male	59	IIA	deceased	relapse	local recurrence and distant metastasis	199	224	unknown
female	58	IB	alive	relapse	local recurrence	1066	1343	yes
female	52	IIB	deceased	progression			263	unknown
male	51	III	deceased	progression			201	unknown
female	63	IIB	deceased	relapse	distant recurrence / metastasis	849	857	yes
male	68	IA	alive	complete remission			1623	yes
female	50	IA	deceased	relapse	local recurrence	379	627	yes
male	53	IB	deceased	relapse	local recurrence	619	1234	no
male	61	IIA	deceased	progression			285	no

142 rows

2017-18: 5 variables → 2018-19: 9 variables
2017-18: 50 patients → 2018-19: 142 patients

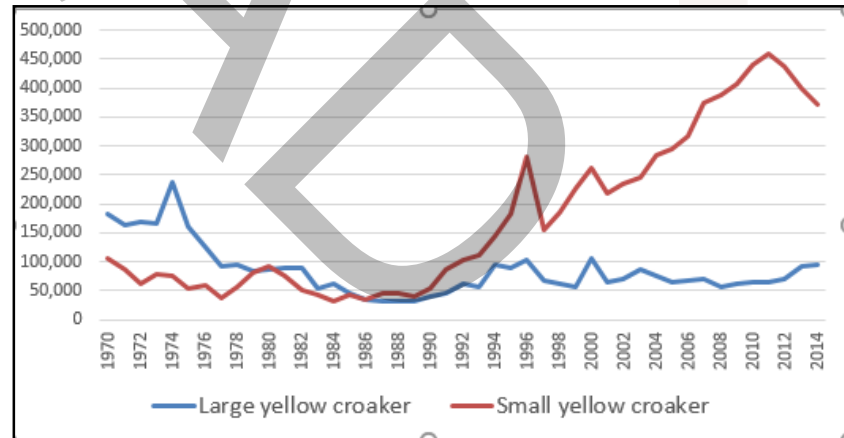


2018-19: FISH CATCH DATA

Marine Species	Annual global capture for the species in the wild (tonnes)									
Common Name	(1t =1000 Kg)									
	1970	1971	1972	1973	1974	1975	1976	1977	#	2014
Large yellow croaker	182,100	163,000	169,300	166,700	237,349	161,163	123,670	#####	#	96,073
Small yellow croaker	104,900	85,400	62,000	77,200	76,030	54,855	58,150	#####	#	370,520
Largehead hairtail	562,911	610,073	735,077	800,783	862,093	697,964	620,861	#####	#	1,260,824
Japanese anchovy	419,500	417,600	473,900	430,900	460,973	420,615	342,866	#####	#	1,396,312
Blue swimming crab	23,600	24,400	27,200	28,000	37,068	32,225	32,667	#####	#	212,571
Southern rough shrimp	4,350	6,238	10,859	7,254	4,899	5,278	6,449	#####	#	320,162
Fleshy prawn	14,814	13,331	12,399	36,164	42,475	34,297	9,596	#####	#	140,936

Possible classifications:
 2 yellow croakers
 2 major food fish
 3 shellfish

graph example:



TASK SHEET GLOSSARY

Glossary of terms in the task data:

- **Pancreatic:** relating to the pancreas
- **Tumour stage (TNM):** TNM is a system of classifying stages of cancer according to seriousness
- **Complete remission:** treatment has removed all the cancer cells (they are not visible on tests)
- **Progression:** the cancer has continued to grow or has spread to other parts of the body
- **Relapse / Recurrence:** both terms mean that cancer has returned after a period of remission
- **Relapse-free interval:** the period of remission from cancer before a recurrence
- **Local recurrence:** the cancer has appeared again in the same part of the body
- **Distant recurrence / Metastasis:** both terms mean that the same type of cancer has appeared in another part of the body (e.g. in the liver or brain)
- **First degree relatives:** parents or siblings (more distant relations are second or third degree relatives)
- **Capture (of fish):** wild fish catch (i.e. not including production from aquaculture / farmed fish)
- **Nutritional data:** the amounts of different nutrients and other substances contained in food (see the recommended websites for examples of what this includes)



2018-19 ADDITIONAL TASKS

Additional formative tasks with tutor feedback:

- **CW1 Formative task:** Prepare three Research Questions (RQs) and three graphs which display the data that answer your RQs.
- **Objectives:** *Students formulate aims before analysing data*
Students design clear and informative graphs
- **CW2 Formative task:** Prepare an annotated bibliography of 5 or more sources that you can use in your literature review.
- **Objectives:** *Check that students format references correctly*
Check that students use sufficient number of academic sources

Secondary objective: *support students' time management*

Additional lesson: creating different graph designs in Excel, additional basic statistics (e.g. correlation coefficient)



OPPORTUNITIES FOR LEARNING

Science report structure

- Coherence between sections (Literature review – RQs – Results – Discussion)

Language Skills, e.g.

- Method – Passive voice
- Discussion – tentative language, modals
- Literature review – reporting language
- Results – describing trends; proportion; comparison
- Technical terminology (topic-related; statistics)

Critical thinking skills:

- RQ design
- Data selection and analysis

Research and academic integrity:

- Finding and evaluating sources
- Source integration and referencing

Time management: task has multiple stages



CHALLENGES

Locating sources of scientific data:

- Science databases = highly technical; may require software
- Data sets published with one article → risk of plagiarism

Data source ideas:

- *Background data in technical databases e.g. DNA samples' biodata*
- *International Organizations: UN / WHO / World Bank*
- *Wildlife NGOs*
- <http://ghdx.healthdata.org/data-sites-we-love>
- Be prepared to repackage data in different formats (e.g. FAO fish catch data)

Engaging Chemistry students:

? *PubChem*: profiles of chemical elements/compounds



CONCLUSION

Using Secondary Data:

- Students can engage with scientific data without time lost for data collection
- Avoids safety / ethical objections
- Multiple learning opportunities

Maximizing effectiveness:

- Providing data in a suitable format
- Strategic scaffolding tasks

Challenges:

- Finding data sources
- engaging students from different science majors



REFERENCES

- Haaker, M. & Morgan-Brett, B. (2017). 'Developing research-led teaching: two cases of practical data reuse in the classroom' *Sage Open* April-June: 1-9. <https://doi.org/10.1177%2F2158244017701800>
- Hox, J.J. & Boeijs, H.R. (2005) 'Data collection, primary vs. secondary'. *Encyclopedia of Social Measurement*, 1: 593-599, Elsevier
- Pienta, A.M., Alter, G.C. & Lyle, J.A. (2010) 'The enduring value of social science research: the use and reuse of primary research data' Paper presented at "The Organization, Economics and Policy of Scientific Research" workshop, Torino, Italy, April, 2010
- Seviour, M. (2015) 'Assessing academic writing on a pre-session EAP course: Designing assessment which supports learning'. *Journal of English for Academic Purposes*, 18, 84-89

Data sources:

- ICGC (International Cancer Genome Consortium), 2017, United States – Lung Cancer, [Online] Available from: <https://dcc.icgc.org/releases>
- ICGC (International Cancer Genome Consortium), 2018, Canada – Pancreatic Cancer, [Online] Available from: <https://dcc.icgc.org/releases>
- BTO (British Trust for Ornithology), 2017, BBS bird population trends [Online] Available from: <https://www.bto.org/volunteer-surveys/bbs/latest-results>
- UN FAO Fisheries and Aquaculture Department, 2018, FAO Fishfinder, [Online] Available from <http://www.fao.org/fishery/species/search/en>

Special thanks to Dr. Eduardo Medina-Roldán (ENV) and Dr. Theo Kantidakis (BIO)



THANK YOU



Xi'an Jiaotong-Liverpool University

西交利物浦大學