New Graduate Course Proposal Form
Faculty of Health

The following information is required for all new course proposals. Provide evidence of consultation, where appropriate. To facilitate the review/approval process, please use the headings below (and omit the italicized explanations below each heading).

All new course proposals must include a library statement.

1. Graduate Program:
Psychology

2. Responsible Unit:
Unit responsible for the course, e.g., Faculty Departments.

Psychology

3. Subject Code ¿Rubric?and Course Number:
PSYC 6120

4. Credit Value:
3.0

5. Long Course Title:
Strategies for Data Management and Data Cleaning

6. Short Course Title:
This is the title that will appear on University documents where space is limited, such as transcripts and lecture schedules. The short course title may be a maximum 40 characters, including punctuation and spaces.

Data management & data cleaning

7. Effective Term{Calendar Year:
Fall{2024

8. Language of Instruction:
English

9. Mode of Delivery:
More than one mode can be designated.

SEMR

10. Calendar Short Course Description:
This is the description of the course as it will appear in the University course repository and related publications. Calendar short course descriptions should be written in the present tense and may be a maximum of 60 words. Please include information with respect to any pre-{co-requisites and{or cross-listing or integration in the course description.

This course will cover the essential steps that precede quantitative data collection and the proper handling of data prior to the central statistical analysis designed to test research hypotheses. Issues related to sampling, data management, visualizing and cleaning data, and preparing data for analysis will be covered. Assignments will focus on developing hands-on practical skills relevant to practicing researchers.

11. Expanded Course Description:
This is the detailed course description that will be published in course outlines, program handbooks, etc. Expand upon the short description in order to give academic approval committees a full and clear sense of the aims and objectives of the course and the types of materials it will cover.

How data is managed as it is collected and how these data are treated before conducting inferential analyses are two key aspects of research. If data is mis-handled in any way before analysis, or if data are not appropriately cleaned, then the results of any hypothesis tests will be severely compromised. As the old adage says, ' garbage in, garbage out.' In this course, we will learn to prevent garbage from entering our final dataset that we analyze to test our key hypotheses. Although the details of what is covered will differ slightly each year based on the class composition—with the techniques covered tailored to suit the research being done by students each year—several core principles will always be covered. Specifically, we will start by discussing issues related to sampling: from whom our data will be collected. Then we will learn to develop a proper data management plan so that there is always a good record of how data was moved and treated at each step ¿e.g., from servers for online survey services (like Qualtrics) to the Open Science Framework?, and how it will be stored and preserved. This is followed by an examination of how data is entered into a spreadsheet and subsequently loaded into software for analysis. The next major step involves understanding our data and ensuring it is of high quality and appropriate for analysis. This means visualizing data to look for potentially problematic patterns,

identifying inattentive responders, excluding participants based on principled criteria ¿e.g., treatment of outliers?, transforming data when necessary, and handling missing data ¿e.g., imputation?. Hands-on assignments using real or generated datasets will be used to illustrate core concepts and practice the techniques required.

12. Course Learning Outcomes:
¿Necessary for Quality Assurance approval and cyclical program reviews?
What will students be able to do upon completion of this course specifically!

Upon completion of this course, students will be able to:

- Explain how sampling affects statistical inferences
- Design a data management plan appropriate for their own research
- Visualize and describe data in a variety of ways in order to discover potential problems
- Discuss various techniques for identifying inattentive responders
- Recommend several different methods for identifying influential outliers
- Explain when data should be transformed and discuss various methods for doing so
- Categorize various forms of missing data and be familiar with some methods for dealing with missingness

13. Rationale:
Please indicate how the proposed course will contribute to the academic objectives of the program. Please provide a description of the learning outcomes{objectives for the course. As well, please indicate the relationship of the proposed course to other existing options, particularly with respect to focus{content{approach. If overlap with other existing courses exists, please indicate the nature and extent of consultation that has taken place. Additionally, please append the graduate program¶s existing learning outcomes as a separate document.

The primary aim of the graduate program is to train psychological scientists, individuals adept at conducting various forms of research in the service of answering psychological questions. A majority of the research conducted in the department, and in Psychology in general, is quantitative in nature. Managing quantitative data and treating it appropriately before analysis is an absolutely fundamental aspect of analyzing and making inferences based on a quantitative approach. The primary objective of this course is to teach graduate students how to manage data from the collection phase all the way to the main analysis in a manner that ensure that these data are trustworthy and appropriate for analysis ¿see Section 12 for specific learning objectives?. However, these fundamental steps are rarely taught in graduate school and are currently not taught in any systematic fashion in our department or within the

Faculty of Health. In response to my proposal for this course, one faculty member from the Quantitative Methods area observed that ' most grad students have poor data management skills.' Similarly, Dr. Cribbie responded with, ' I love the idea ... it is a huge loophole in our grad student training. I can see this as a really valuable course in the first semester of their MA program.' It appears that there is definitely a need for this course in our department. To properly confirm that this is indeed a loophole in our current training, I e-mailed the other departments in the Faculty of Health to enquire whether any similar courses were being offered there. In response, I learned that although several courses touch upon these ideas, none treat it as a focus or cover it in depth. For example, Dr. Michael Rotondi states, ' I can confirm that we do not discuss data management{cleaning in any detail in Univariate KAHS 6010. We do talk a little bit about some aspects of data management in KAHS 5020{KINE 4562 ¿Meta-analysis and Systematic Reviews?in the context of RevMan, but it is minimal - so I have no concerns with potential overlap in any of my courses.' Dr. Alison Macpherson stated, ' I do touch on data management in my Multivariable course, but it is not the focus at all.' With respect to Machine Learning for Health ¿HLTH 6240?, Dr. Christo El Morr explains that data management and data cleaning are ' only a chapter or two in the course and not the main focus of the course' and that ' I think your proposed course will be an excellent addition.'

In terms of academic requirements, this course will qualify as a Quantitative Methods course, satisfying requirements for statistics courses and helping to satisfy the Quantitative Methods diploma requirements.

This course also reflects the priorities of our federal granting agencies, with the Tri-Agency Research Data Management policy making it clear that data management plans are essential for describing how data is ' collected, documented, formatted, protected and preserved.'

14. Evaluation:

Please supply a detailed breakdown of course requirements, including the type and percentage value of each assignment. The expectation is that course assignments can normally be accomplished within the course period. If applicable, details regarding expectations and corresponding grading requirements with respect to attendance and participation should be provided.

Take-Home Assignment 1 ¿10⁄ ?: Sampling and Sample Size

Conduct sample-size estimation for a proposed study or a published study lacking such information. This may be an a priori power analysis ¿rooted in null-hypothesis statistical testing?or a precision analysis.

Take-Home Assignment 2 ¿10⁄ ?: Entering and Organizing Data
Consistent with your own type of research, demonstrate at least 2 different ways in which data can be collected and entered into a spreadsheet file, at least 3 different ways of importing data of various types into your analysis software, and at demonstrate a facility for performing the following operations: re-naming variables, deleting variables, deleting cases, and identifying cases that satisfy different conditions ¿e.g., scores greater than X?.

Take-Home Assignment 3 ¿10⁄ ?: Identifying Problematic Patterns in Data
Visualize an entire dataset to identify potential patterns and problems that will need to be resolved ¿e.g., patterns of missingness, incorrect variable types?.

In-Class Presentation ¿10⁄ ?: Creating a Data Management Plan
Create and orally present a data management plan appropriate for your own research. The plan should describe how data will be collected, the documentation that will accompany these data ¿e.g., codebook?, the format of the data at each step of the process, how{if the data will be moved and stored at each step, how the data will be protected, and lastly how the data will be preserved and possibly shared ¿e.g., data deposition?.

Take-Home Assignment 4 ¿10⁄ ?: Identifying and Accounting for Inattentive Responding
Using an actual dataset, identify inattentive responders and propose a plan for dealing with these participants.

Take-Home Assignment 5 ¿13⁄ ?: Identifying and Managing Outliers
Using an actual dataset, explore at minimum 3 different methods for identifying outliers and compare{contrast the results. Propose a plan for how to deal with these outliers.

Take-Home Assignment 6 ¿14⁄ ?: Transforming Data
Using an actual dataset, identify data that requires transformation, transform that data in a minimum of 3 ways, and report how those transformations affect a set of analyses.

Take-Home Assignment 7 ¿14⁄ ?: Accounting for Missing Data
Based on a dataset, identify missing data and categorize the type of missingness observed. Examine at minimum 2 different ways of dealing with this missingness and weigh their strengths and weaknesses.

Participation and Attendance ¿10⁄ ?:
Attendance for in-class lectures and participation in the discussion of class topics, along with sharing of resources for these topics, will all inform the participation grade.

15. Integrated Courses:
Graduate courses may be integrated only with undergraduate courses at the 4000-level, where it is understood that 4000-level indicates an advanced level. Graduate students will be expected to do work at a higher level than undergraduates. If the proposed course is to be integrated, please provide a grading scheme that clearly differentiates between the work that undergraduate and graduate students perform, including a description of how the work performed by graduate students is at a higher level. As well, please indicate the course information for the undergraduate course ¿i.e., Faculty{unit{course number{credit value?and include a statement from the relevant undergraduate chair or undergraduate director indicating agreement to the integration.

N{A

16. Cross-listed Courses:
Cross-listed courses are offered between two or more graduate programs. For cross-listed courses, please include a statement of agreement from the director of the other graduate program¿s?

N{A

17. Enrolment Notes:
Is the course limited to a specific group of students; closed to a specific group of students; and{or if there is any additional information necessary for the student to know before enrolling.

N{A

18. Faculty Resources:
Provide the names of faculty members in your program qualified to teach this course. Stipulate the frequency with which you expect this course to be offered, including the impact that this course will have on faculty resources.

Qualified Faculty members: Chris Green, Michael Friendly, Monique Herbert, Raymond Mar, Robert Cribbie

This course is anticipated to be taught every two years or so. I believe it will be easy for me to incorporate this course into my teaching schedule without major disruption and so this will course will have little impact on faculty resources. Especially as the responsibility for teaching this course can be spread over the faculty listed above.

19. Physical Resources:
Please provide a statement regarding the adequacy of physical resources ¿equipment, space, labs, etc.?, including whether or not additional{other physical resources are required and how the need for these additional{other physical resources will be met.

No physical resources will be required, aside from a classroom to host the course. Students will use their own computers for the completion of assignments and the focus will be on software and tools that are free and open-source.

20. Bibliography and Library Statement:
Please provide an appropriate and up-to-date bibliography in standard format. A statement from the University librarian responsible for the subject area certifying that adequate library resources are available for the new course must be provided.

Aggarwal, C. C. ¿2017?. An introduction to outlier analysis. In *Outlier analysis* ¿pp. 1-34?. Springer, Cham.

Boukerche, A., Zheng, L., & Alfandi, O. ¿2020?. Outlier detection: Methods, models, and classification. *ACM Computing Surveys ¿CSUR?, 53¿3?*, 1-37.

Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. ¿2016, June?. Data cleaning: Overview and emerging challenges. In *Proceedings of the 2016 international conference on management of data* ¿pp. 2201-2206?.

Dasu, T., & Johnson, T. ¿2003?. *Exploratory data mining and data cleaning*. John Wiley & Sons.

Ganti, V., & Sarma, A. D. ¿2013?. Data cleaning: A practical perspective. *Synthesis Lectures on Data Management, 5*¿3?, 1-85.

Hodge, V., & Austin, J. ¿2004?. A survey of outlier detection methodologies. *Artificial intelligence review, 22*¿2?, 85-126.

Ilyas, I. F., & Chu, X. ¿2019?. *Data cleaning*. San Rafael, CA: Morgan & Claypool.

Mertz, D. ¿2021?. *Cleaning Data for Effective Data Science: Doing the other 80⁄ of the work with Python, R, and command-line tools*. Packt Publishing.

Osborne, J. W. ¿2013?. *Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data*. Sage.

Rahm, E., & Do, H. H. ¿2000?. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull., 23*¿4?, 3-13.

Rousseeuw, P. J., & Hubert, M. ¿2011?. Robust statistics for outlier detection. *Wiley interdisciplinary reviews: Data mining and knowledge discovery, 1*¿1?, 73-79.

Singh, K., & Upadhyaya, S. ¿2012?. Outlier detection: applications and techniques. *International Journal of Computer Science Issues ¿IJCSI?, 9*¿1?, 307.

Van den Broeck, J., Argeseanu Cunningham, S., Eeckels, R., & Herbst, K. ¿2005?. Data cleaning: detecting, diagnosing, and editing data abnormalities. *PLoS medicine, 2*¿10?, e267.

Van der Loo, M., & De Jonge, E. ¿2018?. *Statistical data cleaning with applications in R*. John Wiley & Sons.

Please submit completed forms and required supporting documentation by email to Pina Guzzo-Foliaro, Administrative Secretary Research – pdimaria@yorku.ca

YORK
UNIVERSITÉ
UNIVERSITY

YORK UNIVERSITY LIBRARIES

203L Scott Library
4700 Keele St.
Toronto ON
Canada M3J 1P3
www.library.yorku.ca/

Memo

To: Professor Raymond A. Mar, Psychology Department, Keele Campus

From: Priscilla Carmini, Scholarly Communications Librarian, Scott Library

Date: 31 August 2022

Subject: Library Support Statement for PSYC 6120: *Strategies for Data Management and Data Cleaning*

I have reviewed the course proposal material for *Strategies for Data Management and Data Cleaning.* I am happy to report that York University Libraries (YUL) will be able to support this course. Most of the titles in the bibliography are already held at York. Any required titles not held at York will be ordered in a timely manner.

A quick search of York University Libraries resources revealed more sources related to data management and data cleaning. This includes both journals and monographs in print and electronic format.

For further research, students can use OMNI the online catalogue, and periodical indexes such as PsycINFO, Medline, Web of Science, and PubMed. More resources can be found on YUL's Psychology Research Guide and Data Visualization Research Guide.

Amongst other services, students have access to subject specialists for research support and the Resource Sharing Department to request materials not held at YUL.