Title	Knowledge of prior work and soundness of project: A review of the research on secondary analysis of research data
Author(s)	Dennis Kwek and Galyna Kogut
Published by	National Institute of Education (Singapore)

Citation: Kwek, D. & Kogut, G. (2015). Knowledge of prior work and soundness of project: A review of the research on secondary analysis of research data. Technical Report, Office of Education Research. Singapore: National Institute of Education.

Knowledge of Prior Work and Soundness of Project: A Review of the Research on Secondary Analysis of Research Data

By

Dennis Kwek
Galyna Kogut

National Institute of Education
Singapore

November 2015

Introduction

It is a truism of educational and social research that almost all data are seriously under-analysed; unless the data collection is tightly designed to test a specific hypothesis, the original researcher will explore only a fraction of its potential. To attempt to extend the utility of research data, an increasing number of researchers and research institutes are encouraging the archiving and sharing of research data. Subsequently, data, once treated as the private property of individual researchers, are now becoming increasingly available to local and international research communities (Musgrave & Ryssevik, 1999). In other words, they are becoming valuable commodities beyond the lifespan of research projects. This trend has been fuelled by the need to collaborate and share diminishing resources and it has been enabled by the creation of new technologies. In the West, digital archives are increasingly becoming a reality of social science research. Today's researchers are increasingly encouraged to locate, access and analyse data from data archives worldwide, utilising a range of documentation initiatives software systems and network tools.

The deposition of social science data into an archive itself is not novel; indeed, quantitative researchers have long archived their data to maximise the dissemination of research findings. What is new is the pressure by funding bodies on qualitative researchers to archive their data. A number of funding bodies in the UK and USA have made research funding conditional on the depositing of data in a digital archive or repository (Corti, 2000; Parry & Mauthner, 2004). For example, the Economic and Social Research Council (ESRC), a major funding source for UK social search research, requires all award-holders to offer their data for archiving, with this requirement stipulated for both quantitative and qualitative datasets. Failure to meet the requirement can incur a financial penalty and it is within the Research Council's rights to withhold the final instalment of research funding to the holding institution, should this be the case (ESRC, 2003). Although the ESRC is arguably the only research council that has made archiving a condition of research funding, there is a strong suggestion that the other research councils will follow suit along with the larger charity organisations. Some foundations, including the Leverhulme Trust, currently recommend archiving of data (Perry & Mauthner, 2004).

The motivations claimed for requiring the deposit of material within an archive are scientific and pragmatic. Broadly, the justifications put forward are that the publication of data will contribute to a more rigorous scientific approach by facilitating the transparency of research, allowing for the testing and checking of results and interpretations by other researchers. To ensure an ethically responsible approach to using another researcher's data, the deposit of, and access to, archived data is generally subject to rigorous codes of practice. International data archives have strict rules concerning 'responsible use' of the material which are premised on the understanding that intentional identification or disclosure of a person or establishment violates the assurances of confidentiality given to providers of the information (Inter-University Consortium for Political and Social Research, 2012; South African Data Archive, 2013). Archive users traditionally access these datasets solely for statistical analysis and reporting of aggregated information and not for the investigation of specific individuals or organisations. While the vast majority of data archives service the needs of quantitative researchers, there are some exceptions (Corti, 2000). These are the UK Data Archive (2013) and the Murray Research

Archive (2013) in the US, which have been acquiring qualitative data for some years, and the Swiss Information and Data Archive Service (SIDOS, 2013) which has recently begun to acquire qualitative data. In addition, the Finnish Social Science Data Archive (2013) and Danish Data Archive (2013) are currently considering acquiring qualitative data. The Australian Research Council (ARC) has a vested interest in the trajectory towards archiving given that it funds a large proportion of research undertaken in the country. In the UK, the ESRC Qualitative Data Archival Resource Centre (Qualidata) was established in the early 1990s with the aim of increasing the number of qualitative data collections and raising visibility and access to these resources (Qualidata, 2013). In particular, the work of Qualidata has contributed to and helped to fuel a growing international debate over the possibilities and problems of reusing qualitative data and, in so doing, sought to stimulate a culture of data archiving and reuse in qualitative research in the social sciences (Backhouse, 2002; Corti, 1999, 2000; Corti et al., 1995; Thompson 2000).

Equally important is the pragmatic reason to avoid the repetition of expensive research by allowing research that has already been carried out to be exploited as fully as possible. This is achieved by secondary use of data (Dale et al. 1988; Glaser, 1962; Gorard, 2002; Hyman, 1972). As Glaser points out, secondary analysis is seemingly perfectly suited to "the research needs of persons with macro-interest and micro-resources" (Glaser, 1963, p. 11). There is however more to secondary analysis than "easy access to other people's data for the lazy or impoverished researcher" (Smith, 2008, p. 332). Before proceeding into describing the nature of secondary analysis of data and the potentials and challenges of using such data, it is important to provide the Singapore context and what can be perceived to be an increasing demand to ensure that data is archived and prior to that, properly managed. This is followed by a framework for understanding the research data lifecycle. The principles, functions and modes of secondary analysis are described next and this review continues with the potentials and pitfalls of secondary analysis and data sharing, before concluding with recommendations.

The Singapore Context: Current Practices and Change Drivers

Three drivers of change can be seen to be placing pressure on the National Institute of Education (NIE) to consider long term solutions to data archiving of research data. The current practice of 'data archiving' is archaic in relation to established practices internationally and locally (such as Singapore Management University and National University of Singapore): upon project closure, data is stored in physical storage medium such as boxes (for physical artefacts), tapes, CD-ROMs or portable hard disks (for digital data). While a rudimentary file organisational structure is recommended by the Office of Educational Research (OER), no other documentation is necessary (see Figure 1).

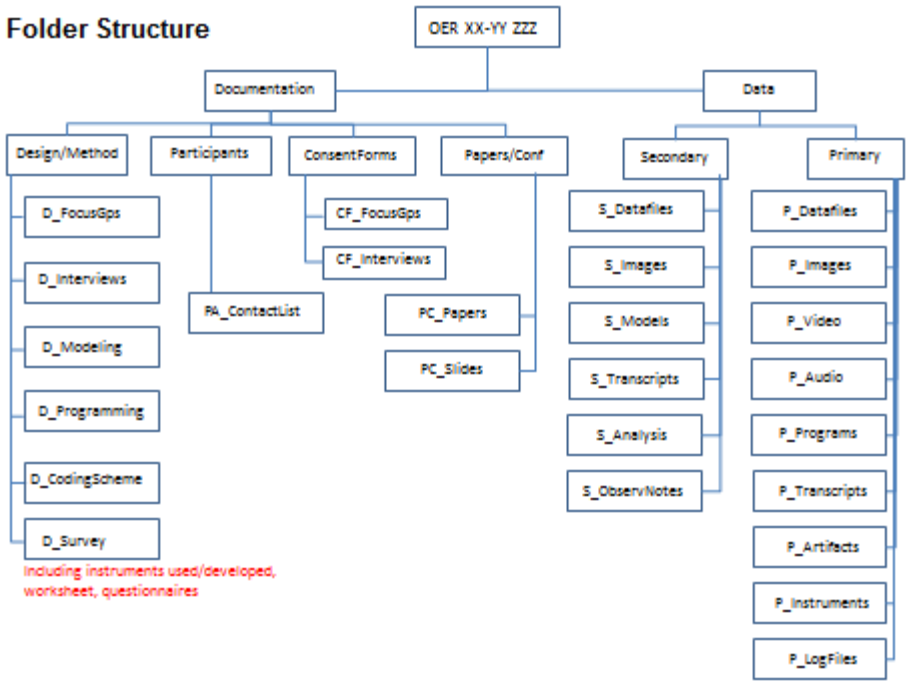

Figure 1: Recommended Folder Structure for Data Archiving (OER)

This means that projects are not required to supply documentations on methodological and sampling considerations for their data, or a detailed index of the contents of the digital data that is 'archived', much less the nature of the content. Instead, it is typical of project researchers to do a 'data dump' of all their research data from their computers into the portable harddisks, with the assumption that it is meant purely for storage. Crucially, there is an implicit assumption that data, once stored, can be discarded or will never be reused in any way. It is not uncommon for data stored in portable harddisks to go missing or become damaged over time. This not only represents a significant loss of funding dollars that were pumped into data collection, it also means that the research potential for such collected data cannot be fully utilised.

This current practice of 'data archiving' is untenable for three reasons. Firstly, the NIE's *Strategic Roadmap 2013-2017* has a component that details the Research, Development and Innovation (RDI) plan which includes "developing NIE as a knowledge hub for policy and practice in teacher education". The Roadmap also suggests developing more cost efficient forms of research work, including sustainable cost reductions through pooling of resources and generating economies of scale, reducing wastage and duplication, and increasing collaborations between research projects. All these can be achieved through the short-term redesigning of data management and archiving practices with a view towards a long-term development of an NIE-wide data management and archiving system. A data archive would reduce wastage and duplication as researchers can reuse secondary data, increase collaborations between the original researchers that collected the data and new researchers keen to access such data, and in the longer term establish NIE as a knowledge hub as the archive is increasingly used for both research and teaching purposes.

Secondly, in September 2010, the Nanyang Technological University (NTU) – of which NIE is a part - has signed an accord on research work in Singapore known as the *Singapore Statement on Research Integrity*. Part of the Statement includes an explicit clause with regards to "research records": "Researchers should keep clear, accurate records of all research in ways that will *allow verification and replication of their work by others*". This implies that secondary analysis of research data is to become an established practice, since it is through secondary analysis that "verification and replication" can be done "by others". Furthermore, the 2012 NTU *Policy on Research Integrity and the Responsible Conduct of Research* states that researchers are required "to *maintain full and accurate records of research and their storage in NTU*, both in hard copy and as electronic records", with such records and outputs "stored and maintained at NTU *for a minimum of 10 years* after publication or patenting". A ten year storage stipulation implies that data cannot simply be stored in volatile hard disks which can get damaged, but a longer-term hardware infrastructure is needed to maintain and store such data in NIE.

Other documents NTU has to comply with on the national level include a recently enacted Singapore Personal Data Protection Act (2013) which intends to protect the rights of individuals to their own data and regulates all activities of an organization aimed at collecting, usage and disclosure of personal data, research data being part of it (individuals can be identified from the research data). This document has to be implemented by all Singapore organizations except the public sector.

In addition to this, there is a set of documents stipulating the procedures for conducting research, data collection, keeping research records, ethics regulations and procedures underpinning NTU/NIE research projects implementation and researchers' behaviour. These include NTU IRB, Research Integrity Framework documents, NTU IP policy, etc., which have been there in action for the past few years but might not have been revised and updated by NTU/NIE. On the other hand it is a newly adopted PDPA, the primary objective of which is to protect the existing data or future data from misuse, violation of privacy, etc. There might be some overlaps or even possible conflicts between the existing and newly adopted policies which need to be taken into account by NTU/NIE in their implementation (i.e. conflicting clauses in PDPA and NTU IRB and Research Integrity Framework related to the procedures for keeping research records and destroying personal data) which need to be solved in order to achieve a smooth implementation of all relevant policies.

To comply with the recently adopted legislation and to bring its implementation in the alignment with the already existing IRB and Research Integrity Policy NIE ACIS unit has taken charge and started work on implementing PDPA requirements. NIE has been working on introducing to relevant units the documents explaining the process of PDPA implementation, such as *Advisory Guidelines for the Education Sector, Advisory Guidelines on the Personal Data Protection Ant for Selected Topics, Closing Note for Public consultation on Proposed Advisory Guidelines for the Education Sector, Singapore's Personal Data Protection Act 2012 Mini Booklet*, developed by PDPC (Personal Data Protection Committee) of Singapore. NIE has also worked on the revisions of the existing documents such as OER RC 9 Form (which is a request for the use of data or access to reports), OER Research Publication Protocols (which provide guidelines for research integrity including publication guideline), ICT Requirements for Research in NIE as

well as drafting NIE Research Data Management Plan as preparation for more efficient data management in future.

Finally, other Singaporean higher education institutes such as the National University of Singapore and the Singapore Management University (SMU) have various institutional policies in place for research data management across all of their research work. SMU in particular has established a research data management service at the Li Ka Shing Library for its social science datasets and to facilitate data sharing across SMU researchers. The SMU data services also include guidelines for research projects on research data management planning, data organisation and documentation, data sharing, data security, data analysis and visualisation. These universities are forward-thinking in establishing guidelines for the management of research data, as well as archiving of data to facilitate secondary analysis or access.

Even if the first and third pressures listed above are downplayed, and even if international movements towards data archiving of research data are ignored, there is, at the very least, a clear legal requirement under NTU's policies to seriously consider the implementation of data archiving processes and a data archiving infrastructure that can attend to the policy stipulations.

A Framework of the Research Data Lifecycle

To better understand the nature of research data, it is useful to provide a definition and a framework that describes how research data is transformed, utilised, maintained and archived throughout a research project. While the term "research data" is often construed to mean only numeric data sets, it can also comprise survey data, collections of images, analyses of texts, and more. The Canadian Association of Research Libraries offers this definition:

> Research data are defined here as the factual records (e.g. microarray, numerical and textual records, images and sounds, etc.) used as primary sources for research, and that are commonly accepted in the research community as necessary to validate research findings (CARL Data Management Sub-Committee, 2009, p. 4).

This suggests that research data refers to the 'raw' data that researchers use or analyse, and that the analytical data such as statistical calculations, coding, tagging, and so on, are not deemed to be research data. If defined this way, there are implications on what data should be stored in a data archive: primary sources of data may be accepted, but derived sources of data may not.

Data itself often have a longer lifespan than the research project that creates them. Researchers may continue to work on data after funding has ceased, follow-up projects may analyse or add to the data, and data may be reused by other researchers. Indeed, to facilitate further use of data, it is imperative that data are well-organised, well-documented and preserved so that they can be invaluable to advance scientific inquiry and to increase opportunities for learning and innovation. The UK Data Archive suggests that data can be seen to flow through various stages in a research lifecycle:

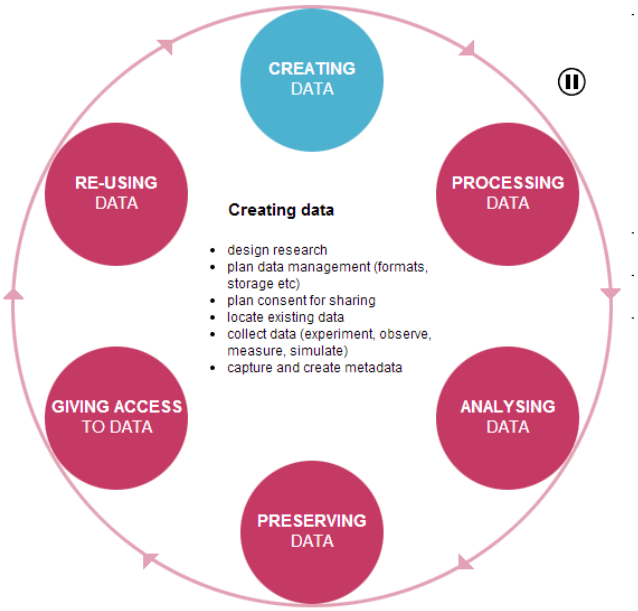

Figure 2: Research Data Lifecycle (UK Data Archive, 2013)

The stages of the Research Data Lifecycle are as follows:

1. *Creating Data*: At this stage, research designs and data management plans are drawn up, consent matters are established, attempts are made to either locate existing data or to collect data, and importantly, plans are made to capture and create metadata to make the data meaningful for future use.
2. *Processing Data*: Data are then transformed – digitised, transcribed, translated, entered into data analytical software. They can be cleaned, verified, checked, anonymised, and stored in an organised manner.
3. *Analysing Data*: Data are interpreted, derived, with the intention to produce research outputs and publications. Data are then prepared for preservation.
4. *Preserving Data*: Once the project has completed using the data, they can be migrated to a suitable medium in an appropriate format, backed up and stored. Metadata and documentation to enable searching and indexing are created.
5. *Giving Access to Data*: Data once archived can be opened up for sharing and distribution. Control access are established to limit use, copyright issues are established to determine who owns the data or what authorship rules should be adhered to. The availability of the data can be promoted to encourage access.
6. *Reusing Data*: Data can be used for further follow-up research or new research. They can be used to undertake research reviews, to scrutinise or verify findings, or for teaching and learning purposes in the form of training materials or examination materials.

The lifecycle is cyclical in that once data is reused, they can lead researchers to form new research ideas, proposals, research questions, directions that subsequently lead to the

generation of new data. What is important about this lifecycle is that for NIE's current practices, the cycle stops at Stage 4 and even then, in a limited manner – data is merely stored with a minimal folder organisation and with poor to no documentation. There is a need to recognise Stages 5 and 6 and how they can contribute to the creation of new data. It is to the secondary access, reuse and analysis of data that we now turn to.

Secondary Analysis of Data

Secondary analysis is best known as a methodology for doing research using pre-existing statistical data (Broom, Cheshire & Emmison, 2009; Gilbert, 1993; Heaton, 2004; Smith, 2008). Social scientists in USA and Europe have been making use of such data throughout the twentieth century, although it was not until 1972 that the first major text on the research strategy, *Secondary Analysis of Sample Surveys: Principles, Procedures and Potentialities* by Herbert Hyman, was published. Since then, the literature on secondary analysis of quantitative data has grown considerably as the availability and use of these data have expanded. There is now a substantial body of work exploring different aspects of the methodology, including several textbooks describing the availability of statistical data sets and how they can be used for secondary research purposes (Dale et al., 1988; Hakim 1982; Kiecolt & Nathan, 1985; Steward & Kamins, 1993) as well as critical commentaries on the scientific, ethical and legal aspects of sharing such data in the social sciences (Fienberg et al., 1985; Hedrick, 1988; Sieber, 1988, 1991; Stanley & Stanley 1988). Accordingly, the terms 'secondary analysis', 'secondary data' and 'data sharing' have become synonymous with the reuse of statistical data sets.

However, in recent years interest has grown in the possibility of reusing data from qualitative studies. Since the mid-1990s, a number of publications have appeared on the topic written by researchers who have carried out ground-breaking secondary analysis of qualitative data (Heaton, 1998; Hinds et al., 1997; Mauthner et al., 1998; Szabo & Strang, 1997; Thompson, 2000; Thorne, 1994, 1998), by archivists involved in the preservation of qualitative datasets for possible secondary analysis (Corti et al., 1995; Corti & Thomson, 1998; Fink, 2000; James & Sorenson, 2000), and by academics interested in these developments (Alderson, 1998; Hemmersley, 1997; Hood-Williams & Harrison, 1998). The extension of secondary analysis to qualitative data raises a number of questions about the nature of this research strategy with some arguing that research data derived from interpretive approaches in the social sciences typically involve subjectivities and epistemologies that do not lend themselves to data archiving (Hammersley, 1997; Parry & Mauthner, 2004). While highly differentiated between researchers and disciplinary backgrounds, the practice of qualitative research is generally seen as one of 'generating' rather than 'collecting' data, with data being co-produced by the researcher and the research participants (Moore, 2007). In this sense, the role of the researcher is foregrounded rather than eliminated, and the idea that data can be neutralized and deposited into an archive, ready to be 'picked up' by others, sits uncomfortably for many. Nevertheless, the establishment of qualitative data archives such as Qualidata suggests that despite epistemological and methodological concerns, some qualitative researchers have found that it is possible to conduct secondary analysis of qualitative data, within certain parameters, constraints and limits. We will describe these below when we discuss the potentials and pitfalls of secondary analysis of data, including qualitative data.

Principles of Secondary Analysis

As mentioned, a key principle of secondary analysis is that it involves the use of pre-existing data. This is exemplified by statements such as “secondary analysis must, by definition, be an empirical exercise carried out on data that has already been gathered or compiled in some way” (Dale et al., 1988, p. 3). There are however significant differences in the types of pre-existing data which can be subjected to secondary analysis, depending on the nature and origin of the material. Such data can be quantitative, longitudinal or qualitative in nature, with each elaborated below.

Quantitative Data

Pre-existing data used in quantitative secondary analysis has been derived from various activities, including research projects carried out by academics, government agencies and commercial groups, as well as the administrative work of public authorities and other organisations that routinely keep records for management purposes. For example, early social researchers often used census and administrative records; Durkheim (1952) used both types of data in his classic study on suicide. Hakim (1982) distinguishes multipurpose – *omnibus* – surveys from those designed with an exclusive primary focus. Omnibus surveys are therefore carried out in order to provide data for multiple uses and users. In contrast, more exclusive surveys are designed to investigate particular research questions and are conducted on a one-off or less regular basis. Given that such surveys are designed to examine specific issues, the scope for secondary analysis may be more restricted (although not necessarily useless) compared to data derived from omnibus social surveys (Dale et al., 1988, p.9). As an example, the Core 2 Panel 2 teacher and student surveys address particular research questions about the nature of teaching and learning in Singapore classrooms. However, it is nevertheless possible to investigate other research questions using the same survey data. Some examples of the items in the survey are included in the Appendix Three below.

Longitudinal Data

Some statistical datasets have been generated for the purposes of both primary and secondary research uses. For instance, longitudinal studies follow up a population cohort over time, collecting data on topics of long-term interest and further topics for investigation may be introduced as the study progresses. These data are subjected to primary analysis by the researchers who collected the data to address particular research questions, but importantly, they may also be archived and used by other (or the same) researchers to address additional or recurring research questions. For example, it is the plan of Core 3 to develop a longitudinal dataset with the Core 2 dataset as the original baseline, and the new Core 3 dataset as a longitudinal dataset that can be used for comparison of changes in teaching and learning over time, in this case a period of approximately 5 years.

Qualitative Data

Importantly, the Core 2 dataset contains not just quantitative survey data, but qualitative data in the form of video and audio recordings and interviews. Such pre-existing qualitative data can be

categorised broadly into *naturalistic data* and *non-naturalistic data* (Heaton, 2004). Non-naturalistic data are solicited by researchers, while naturalistic data are 'found' or collected with minimal structuring by researchers.

Type	Examples	Methodology
Non-naturalistic or artefactual data (solicited for research studies)	Field notes, observational records, interviews, focus groups, questionnaires (responses to open ended questions), diaries (solicited), life stories	Secondary analysis
Naturalistic data (found or collected with minimal interference by researchers)	Life stories, autobiographies, diaries (found), letters, official documents, photographs, film, social interaction	Documentary analysis Conversation analysis

Table 1: Non-naturalistic and Naturalistic Qualitative Data

In qualitative research, Heaton (2004) argues that secondary analysis is conceptualised as a methodology for the study of non-naturalistic or artefactual data derived from previous studies, such as fieldnotes, observational records, and tapes and transcripts of interviews and focus groups. However, unlike quantitative research, there is no tradition of reusing data from previous qualitative studies. By contrast, naturalistic data such as diaries, essays and notes, autobiographies, dreams and self-observation, photographs and film have traditionally been studied using documentary analysis (Jupp & Norris, 1993; Plummer, 1983, 2001; Scott, 1990).

An interesting methodology which has been used in educational and social research to study naturalistic data in the form of recordings of everyday social interaction is conversation analysis (Atkinson & Heritage 1984; tenHave, 1999). These data include tapes and transcripts of audio and/or visual recordings made and transcribed by the researchers personally, or by other researchers. Indeed, conversation analysts often develop analyses based on pre-existing data that have been shared within this disciplinary network. Lengthy extracts of the data, transcribed using a detailed system of annotating verbal and non-verbal interaction, are also reproduced in published studies in order to allow for independent access to, and scrutiny of, the data upon which the analyses are based. However, despite these practices, no distinction is made between 'primary' and 'secondary' conversation analysis. Rather, these naturalistic data are seen as pristine and in no way an artefact of the research process; as such, they are assumed to be open to analysis by all on an equal basis, regardless of who collected the data and rendered it for analysis. The 'naturalistic' status of transcripts have come under critique by some researchers who argue that the process of transcription itself renders the data into a highly interpretative, and subjective, form (Edwards, 2001; Ochs, 1979).

Functions of Secondary Analysis

Three broad functions of secondary analysis can be discerned in the literature. Secondary analysis can be used to investigate new or additional research questions, to verify, refute or refine existing research, or to synthesise research.

Investigation of New or Additional Research Questions

As pointed out, secondary analysis allows researchers to use data that were originally collected for other research purposes. Hakin (1982, p. 1) argues that secondary analysis can be seen as "any further analysis of an existing dataset which presents interpretations, conclusions, or knowledge additional to, or different from, those presented in the first report on the inquiry as a whole and its main results". Secondary analysis can therefore involve "the use of an existing dataset to find answers to a research question that differs from the question asked in the original or primary study" (Hinds et al, 1997, p. 408). Essentially, secondary analysis can be conceptualised as a methodology for conducting independent studies using pre-existing data originally collected for other purposes. A point to note is that data from longitudinal studies (such as the envisaged Core 3 design) is designed *both* to address particular primary research questions and to supply long term data for future secondary research. Examples of research questions from secondary analysis projects using the Core 2 data can be found in Appendix Two, along with the original research questions of the Core 2 project in Appendix One.

Verification, Refutation and Refinement of Existing Research

Secondary analysis can also be used as a means of verifying, refuting or refining the findings of primary studies through the re-analysis of datasets. In this function, secondary analysis involves the "re-analysis of data for the purpose of answering the original research question with better statistical techniques, or answering new questions with old data" (Glass, 1976, p. 3). McArt and McDougal (1985, p. 54) elaborates that secondary analysis involves the

> Use of data gathered for the purposes of a primary research analysis (original research) but looks at questions not addressed by the original investigator, or addresses the same questions using different methods of analysis. In large-scale studies, secondary analysis is further used by investigators to validate the results from the primary analysis; that is, re-analysis and testing of new hypothesis may support further or dispel initial findings.

The re-analysis of data has the potential to generate additional knowledge and insights through the production of findings which may or may not corroborate the primary work. Additionally, where such studies are carried out by independent researchers, re-analyses are also independent of the original research in the sense that they are carried out by other analysts, even though each study investigates the same questions using the same data.

Synthesis of Research

The third, more contentious, function of secondary analysis comes in the form of different types of 'meta-research' designed to synthesise knowledge arising from existing studies. Meta-analysis and related techniques for examining evidence from quantitative research have been described by some researchers as forms of secondary analysis. Others have argued that meta-research strategies are distinct from secondary analysis in that they seek to identify, appraise and aggregate or synthesise existing knowledge on a particular topic. For example, Glass distinguishes meta-analysis from secondary analysis, describing the former as the "analysis of

analyses" (1976, p. 3). Kielcolt and Nathan also claim that meta-analysis is unique in that it seeks to integrate the findings from a 'universe' or 'sample' of investigations of some phenomenon: "that is, the study itself becomes the unit of analysis" (1985, p. 10).

Furthermore, Thorne (1998, p. 548) suggests that "whereas meta-analysis serves as a strategy for synthesis of research findings, secondary analysis provides a mechanism for extending the contexts in which we are able to use and interpret qualitative research data". In addition, Heaton (1998) argues that meta-analysis, meta-synthesis, systematic reviews and literature reviews differ from secondary analysis on two grounds. Firstly, the former research strategies are concerned with appraising and summarising existing knowledge, and not with exploring new research questions or verifying the results of individual studies. Secondly, they mainly involve the study of research reports, seldom reverting to the *raw data* itself.

Modes of Secondary Analysis

Heaton (2004) suggests that there are three main modes of secondary analysis:

1. *Formal data sharing*: Secondary analysis is carried out using datasets that have been *officially* made available for data sharing, with access typically restricted or controlled. Such datasets are usually deposited in general data archives, or those managed by commercial companies, or in the form of raw data published in research reports and other media. In this mode, data is collected and deposited by one group of researchers or organisations - *data donors* - and accessed and used by another group - *data users*. Any secondary analysis using these resources is seen to be independent.
2. *Informal data sharing*: Here, data is either obtained directly from primary researchers and organisations by request, or indirectly through private disciplinary networks (such as researchers in conversational analysis). In this mode, the secondary researchers may or may not invite the primary researchers who donated the data to be part of the research team carrying out the secondary research. Two or more primary researchers may also combine or pool datasets from their previous work and jointly analyse them as part of a new secondary project.
3. *Personal secondary analysis*: This mode does not involve any data sharing. Instead, researchers reuse their own data. Personal or inside secondary analysis is unique in that it is carried out by the same researchers and organisations that originally compiled the data and no one else.

Mode	**Source of data**	**Data donors**
Formal data sharing (via intermediary services)	General data archivers, special collections, commercial companies, public records, publications.	Primary researchers for use in independent secondary studies. Organisations for use in independent secondary analysis.
Informal data sharing (by	Primary researchers' personal	Primary researchers for use in

special request and private networks)	data collections, disciplinary networks, organisations' in-house records.	secondary studies by independent researchers. Primary researches for use in secondary studies by independent researchers and the data donors.
Personal or inside secondary analysis	Primary researchers' personal collections, organisations' in-house records.	Primary researchers for use in personal secondary analyses of their own data. Organisations for secondary use in-house.

Table 2: Modes of Secondary Analysis

A different typology is suggested by Thorne (1994, 1998) which focuses on the nature of the analysis rather than the nature of the sharing. Thorne envisages five discrete types of secondary analysis involving both the reuse of other researchers' datasets and researchers' own data from their previous primary work:

1. *Analytic expansion*: Researchers make further use of their own data "to answer new or extended questions" (Thorne, 1998, p. 548).
2. *Retrospective interpretation*: Researchers examine new questions which were raised but not addressed in the context of the primary study.
3. *Armchair induction*: Researchers apply inductive analytical methods for the purposes of theory development.
4. *Amplified sampling*: This involves the comparison of several distinct and theoretically representative datasets.
5. *Cross-validation*: This involves the use of pre-existing, independently collected datasets, this time in order to "confirm or discount new findings and suggest patterns beyond the scope of the sample in which the researcher personally has been immersed" (Thorne, 1994, p. 267).

To apply this typology, it is suggested that the proposed Core 3 programme will encompass a range of primary and secondary analyses, using primary and secondary datasets. For example, researchers in Core 3 may embark on analytic expansion by reusing the Core 2 video data in the context of new theoretical questions, or Core 3 may compare between Core 2 video data and new video data collected under Core 3, hence involving amplified sampling procedures. The Thorne typology and the Heaton modes are therefore useful heuristics in conceptualising and framing what kinds of secondary analysis a researcher can embark on.

Potentials and Pitfalls of Secondary Analysis and Data Sharing

Debates around data archiving and the reuse of research data have been gathering momentum over the last decade. This is especially the case for qualitative data stemming from the establishment of Qualidata (Bishop, 2005; Hammersley, 1997; Mauthner et al., 1998; Moore, 2007; Parry & Mauthner, 2004). Since then, there has been significant interest in the issues

surrounding the archiving of qualitative, as well as quantitative, data. Such issues revolve around the potentialities of secondary analysis of quantitative and qualitative data, as well as the challenges around reusing such data. The debates are especially polarised in the context of qualitative data. For example, initial concerns in other national contexts where archiving of qualitative data has been implemented have tended to centre on the epistemological issues around sharing data and the notion of qualitative research as an insider activity (Mauthner et al., 1998) that involves interpretation and subjectivities not concrete enough to deposit in an archive. The idea of data as embedded in, or unable to be separated from, the theoretical positioning of the individual researcher is therefore a pertinent critique that has emerged within the social science literature. From this perspective, the way qualitative research is conducted is grounded in underlying theoretical assumptions which are not always explicit. While this has long been acknowledged by qualitative researchers, Hammersley (1997) argues that the idea of a databank reverts back to foundationalist and positivistic assumptions that knowledge and facts are simply lying around ready to be collected, rather than being co-created and value-driven. More recently, Moore (2007) makes a similar criticism of secondary analysis, arguing that it assumes data are pre-existing 'out there' rather than iteratively constructed by the researcher. On the other hand, such criticisms have been rejected by Van den Berg (2005) for overestimating the constructed nature of qualitative research as derivatives of the specific aims and assumptions of the original researcher. Instead, the empirical, he argues, is undoubtedly connected to the theoretical, but it also has a momentum of its own.

Mindful of current ongoing debates, we firmly believe that there are benefits to data archiving and secondary analysis that outweigh the disadvantages, although we must be mindful of these challenges and take them into account whenever any secondary analysis is to be conducted on pre-existing data. As Glaser points out in the early 1960s: "Secondary data analysis can help save time, money, career, degrees, research interests, vitality and talent, self-images and myriads of data from untimely, unnecessary and unfortunate loss" (Glaser, 1963, p. 14).

Benefits of Sharing and Conducting Secondary Analysis of Research Data

A range of benefits can be seen to be generated from the sharing of research data and the secondary analysis that can be conducted on them.

1. *Cost*: For many researchers, the cost of collecting data is prohibitive. The hiring of research assistants to collect new data each time a new project begins, or a new set of research questions are generated, is not a cost-efficient approach to doing research especially if data already exists that can be used to address the research agenda and questions. In the case of survey work, there is almost no possibility for small or junior researchers to obtain funding at the level required to conduct representative surveys. In the case of classroom observations, additional costs are required to account for video or audio recording equipment which are of sufficient quality to record noisy classroom environments. Conducting secondary analysis on data such as classroom video recordings or large-scale surveys can reduce cost considerations tremendously.
2. *Time*: A great deal of time can be saved when data collection is no longer necessary. If the data already exist, the researchers can get to work almost as soon as they generate

their research agenda, instead of having to face months to years of preparation if they are to be responsible for their own data collection.

3. *Maximising Potential*: There is an argument for data to be archived because of their enormous but as yet largely untapped potential for secondary analysis. This is applicable particularly in the case of naturalistic data which can be subjected to different analytical approaches and lenses.
4. *Increasing Impact*: Secondary analysis of existing data can help to increase the impact and visibility of the originating research and the original researchers. It explicitly promotes the research that created the data and its outcomes through acknowledging the source of data as a minimum requirement for secondary analysis. It can also provide a direct credit to the original researchers as a research output in its own right, especially if the secondary researchers collaborate with the original researchers.
5. *Quality of Data*: Most large-scale data, such as survey data or classroom observation data, tend to be of a high quality and standard. Such data have typically been cleaned, validated, prepared and indexed for analysis by the original researchers. For example, according to Harrop (1980), in general the surveys conducted by professional researchers working in government organisations are more likely to be of better quality, larger scale and more representative of a population than the "local and frequently non-random samples that often form the basis of surveys carried out by academic social scientists, with students often used as interviewers" (p. 15).
6. *Building a Collaborative Culture*: Secondary analysis can encourage collaborations between the original researchers and the secondary researchers. The culture of data sharing is increasingly recognised throughout the world as researchers move into an era of collaboration across disciplines, institutions and countries. Data archiving, by enabling generations of researchers to collaborate across time, discipline and geography, is key to developing a culture of collaborative inquiry.
7. *Unobtrusiveness*: Secondary analysis is an unobtrusive research method. This has the social and ethical benefit of not collecting additional data from individuals and protecting their privacy by respecting an individual's right to be left alone "free from searching inquiries about oneself and one's activities" (Bulmer, 1979, p. 4). In effect, making data available respects the time and knowledge of research respondents and participants, even if it is initially embargoed (Cheshire et al., 2009); it certainly reduces respondent or participant burden. This, of course, is of particular benefit for research into sensitive issues and of vulnerable and hard-to-reach groups (Dale et al., 1988; Rew et al., 2000).
8. *Democratising*: Secondary analysis can be a very democratic research method. The availability of low cost, high quality datasets means that secondary analysis can "restrain oligarchy" (Hyman, 1972, p. 9) and ensure that "all researchers have the opportunity for empirical research that has tended to be the privilege of the few" (Hakim, 1982, p. 4). As "it is the costs of data collection that are beyond the scope of the independent researcher, not the costs of data analysis" (Glaser, 1963, p. 12), the very accessibility of the data enables novice and other researchers to retain and develop a degree of independence. By circumventing the data collection process, secondary analysis can enable novice researchers and graduate students to gain valuable experience in undertaking independent research in an area of their own interest, as well as presenting

opportunities to publish and present their findings as independent researchers. In this sense, secondary data analysis has a valuable role in the capacity building of research skills as well as in developing an early career researcher's theoretical and substantive interests (Smith, 2005).

9. *Research Auditing*: The opening up of data to outside access and examination can be seen as a form of research auditing, a necessary safeguard against error and misrepresentation in educational research (Bishop, 2007; Bryant & Wortman, 1978). By maximising transparency and accountability, Lincoln and Guba (1985, p. 318-319) sees data sharing as one of the principal techniques for establishing the objectivity and validity of qualitative findings. The requirement to archive data can be seen as a form of audit: "yet another step [towards] rendering social research more publicly accountable, cost-effective and regulated" (Cheshire et al., 2009, p 249). Some journals now require that datasets be linked in their published papers so that these can be reanalysed or validated; the knowledge that it is possible to do this may increase the confidence placed in the authors' conclusions. As Broom et al. (2009, p. 1175) points out: "Once someone makes public claims they have a moral obligation themselves to ensure that the basis for those claims can be scrutinized at some point".
10. *Teaching and Training*: Educational data archives can be extremely useful for the training and teaching purposes. The most obvious application is that graduate students and early career researchers can be given the opportunity to study the products of genuine research processes reflecting a variety of methodological and theoretical emphases. They can examine first-hand the progress of analysis from raw material to finished argument, and reflect upon the strategies and tactics of data collection, as well as upon the many dead-ends, lost opportunities, and errors in judgement that inevitably occur in the research effort. Alternative analyses of a given data set can be considered and alternative strategies for examining issues can be explored. This is not to suggest that such activities completely replace the more traditional approach of training researchers by sending them out to do original research; secondary analysis might however increase the quality and value of such research. Secondary analysis also has an important role in teaching research methods. When teaching the methodologies of survey design, questionnaires from large-scale surveys can be examined for good or bad practice in question wording, scale construction, question ordering (Sobal, 1981). It is also a useful tool for teaching statistics; students can examine patterns and findings using real data so lending the exercise a degree of relevance. Encouraging students to undertake their own secondary analysis allows them the opportunity to test their hypotheses on good quality, large-scale real data (Cutler, 1978; Dale et al., 1988; Sobal, 1981). Additionally, by encouraging graduate students to adopt secondary analysis for at least part of their dissertation research, ethical issues and concerns regarding access to the field and respecting the confidentiality of respondents are reduced or may be avoided entirely.
11. *Teacher Education*: A less obvious pedagogical use of datasets is in teacher education. For example, future teachers can examine and be led to reflect upon the extensive interviews and detailed observations of schooling that all too often are not included in journal articles because of space limitations and are used only selectively for illustrative

purposes in research monographs. A close examination of how raw data is transformed into findings and recommendations can increase teachers' interest in the research process and show them ways to use and evaluate research reports. Access to secondary data can also allow teachers and teacher educators to look at how teaching and learning occur in many different kinds of schools, subject areas, and students.

12. *Methodological*: Secondary analysis can reveal new methodological insights by reflecting on previously conducted research (Mauthner et al., 1998; Savage, 2005). It provides opportunities for the replication, reanalysis and re-interpretation of existing research. It provides researchers with the opportunities to undertake longitudinal analyses, to research and understand past events and to engage in exploratory work to test new ideas, theories and models of research design. Secondary analysis can also enable triangulation with data from other sources, for example, by comparing survey results from different research projects. Such analysis can also reveal serendipitous relationships in the data (Dale et al., 1988).
13. *Theoretical and Epistemological*: Secondary analysis can contribute towards theory development, where according to Hakin (1982, p. 170), it can "allow for greater interaction between theory and empirical data because the transition from theory development to theory testing is more immediate". In removing the lag between research design and analysis, secondary analysis can enable researchers to "think more closely about the theoretical aims and substantive issues of the study rather than the practical and methodological problems of collecting new data" (Hakim, 1982, p. 16). It can also provide a historical perspective through comparing longitudinal datasets (Bishop, 2007; Bornat, 2005; Gillies & Edwards, 2006). Researchers can likewise generate new findings by analysing 'old' data in new contexts or through new theoretical lenses. In fact, secondary researchers may even produce more convincing accounts due to access to wider contextual data, greater resources, including time, the 'wisdom of hindsight', and more sophisticated theoretical frameworks and methods of analysis (Walters, 2009).

Given these benefits, it would appear odd that there are researchers who are against the notion of archiving their data. Aside from the possibility that some researchers feel that the data they collected are their personal property (or intellectual property), the discomfort with archiving among researchers is arguably not universal; for historians, archiving is the only way their discipline can move forward. Furthermore, the differences in perceptions may relate to the extent to which researchers rely on non-naturalistic data (such as interviews) rather than naturalistic data, because for non-naturalistic data, there can arguably be a 'special relationship' between researcher and the research participant, or between researcher and research data, given the personal nature of data production for many researchers. This is especially the case for qualitative research data. However, Mason (2007) and Walters (2009) both argue that having some distance from the data and having a form of emotional detachment can be analytically helpful. As Irwin and Winterton (2011, p 8) explain:

> Primary analysts have a privileged relationship to the data they have generated, but do not necessarily have a privileged claim on the arguments which can be made from those data. Sociological data will support different theoretical understandings, and 'being there' is not the final arbiter of the adequacy of such understandings.

An Example of the Advantages of Secondary Analysis of Video Data

To ground some of the above advantages, it might be worth considering an example of using secondary analysis on video data, of which there are a significant amount of classroom video-recorded data in the Core 2 dataset.

According to Heath (2011), it has long been recognised that the moving image provides extraordinary opportunities for social science research. Video as a visual media seems to provide just the resources that qualitative, as well as quantitative, studies need: it gives the opportunity to capture activities as they arise in natural habitats, such as in the classroom, at home, or in the workplace (Heath, 2011). Video data are therefore often characterised as naturalistic data (Heaton, 2004; Knoblauch et al., 2006; Silverman, 2005). These records can then be analysed repeatedly, and they provide access to fine details of conduct, social organisation, and social interaction, at the very least. Moreover, they can be shared and shown to others, and they provide the opportunity to develop an archive of data that can be subject to a wide range of analytic interests. It also brings new opportunities for credibility and trustworthiness in qualitative research methodology and quantitative (video coding) methodology: video recordings can, for example, be viewed multiple times by multiple people and in some cases even at different times or by different research groups. This makes it easier to subject claims or research findings to debate, or to check the researcher's interpretation against the captured data (Derry et al., 2010). It should nevertheless be noted that videos are artifacts — a document of a certain situation or event (Erickson, 2006; Schnettler & Raab, 2008) — having been recorded for particular purposes and in certain contexts, as well as representing aspects of the recording activity itself (such as camera angles or focus) (Knoblauch et al., 2006). Thus, information derived from video recordings does not give unmediated access to ‘reality’ (Erickson, 2006). As Schnettler and Raab (2008) further point out, to characterise video data as naturalistic data means to recognise both the conservation of a wide range of aspects of a certain event and its construction by the researchers through the means of video technology.

Importantly, sharing video data also means not having to go through the process of gathering new data in each and every research project. From a cost-efficiency perspective, the reuse of video data can be regarded as fruitful for the video research communities, as video studies require both video equipment and time (Szabo & Strang, 1997). It is, however, still a time-consuming process in many ways for both the primary researchers in terms of archiving, and secondary researchers in terms of familiarising themselves with the data (Dalland, 2011). Furthermore, Fielding (2004) emphasises the potential of secondary analysis in avoiding the possibility of certain groups being over-researched. In educational research for example, the reuse of video data unburdens teachers and students by reducing the presence of researchers in schools and classrooms. These aspects of secondary analysis have also been argued with regard to the reuse of quantitative data in educational research (Olsen, 2005).

Challenges and Pitfalls of Secondary Analysis

At a fundamental level, three key assumptions are made of secondary researchers:

- Secondary researchers are professionally competent to carry out their work.

- Secondary researchers have professional integrity in the conduct of their work.
- Secondary researchers universally support protection of anonymity of individual subjects.

It is not common to evaluate or interview secondary analysis users for their competency, integrity or agreement to protect the identities, before permission or access is given to them. While it may not be possible to assess every secondary researcher, the common practice is to bind them with legal agreements as a condition of access. The problem with this approach is that legal clauses become relevant only after the fact; the misdeed or confidentiality breach had already occurred. Yet, this is but one of the many challenges and issues arising from secondary analysis and access. At a practical level, primary researchers need to budget costs and time for anonymising and archiving transcripts and other data if necessary, or supply comprehensive metadata to facilitate reuse. Secondary researchers need to orient themselves to the project through accessing available literature on the project or overviews of data and resources archived by the original researchers. They need to understand the research objectives, design, research questions and methods used for data generation (including interview schedules or other data elicitation tools). They need a grasp of the sample, including knowledge of the sampling decisions and how they relate to the research questions, whether the desired sample was achieved and how it relates to a wider population and/or to theory. In addition, it is useful to understand any implicit as well as planned ways in which the sample was structured. Further contextual information supplied by the originating researchers can help secondary researchers determine the fit between the secondary data and their research objectives (Bishop, 2006; Irwin & Winterton, 2011a, 2011b).

Other challenges revolve around infrastructural, cultural, methodological, epistemological, ethical and legal issues pertaining to secondary analysis. Each will be described in turn below.

Infrastructural Challenges

One of the barriers to reusing data in NIE, and indeed in many countries still, has been the lack of an infrastructure to enable access to the rich research data collected in the academic community. While in some disciplines, professional networks enable high quality data to be shared on an informal basis, the preservation of these sources is not ensured, not does it allow equal access for all researchers. The growing establishment of national data-sharing policies is helping to secure data, provide access and provide support for reuse, although the stock of data archived still rests on the willingness to share by the original researchers. And as mentioned, documenting data to the high standard required to render it accessible can be a huge task, particularly without specific funding to support this. Researchers who have constructed large-scale or perhaps long-running studies may be daunted by the prospect of transforming their data into a usable resource. If research data are to be shared, the infrastructure needs to be in place to offer guidance, support and adequate funding from the start of the project to enable the documentation and archiving of its data. Consequently, researchers see the need for a policy to determine relative levels of investment in data preparation and documentation according to different types of datasets and for different uses. For example, for large-scale or longitudinal studies, the not inconsiderable costs of good data housekeeping, high quality data

documentation, anonymization, and sample maintenance are valued as good investments. With smaller studies, most of the needs for making future archiving possible can also be viewed as useful for the research itself, as a form of structuring and good housekeeping. This requires a shift in researcher mind-set from perceiving data as a private resource to be 'dumped' into storage, to data as a usable resource for archiving and sharing.

Issues related to archiving procedures and storage have to be carefully addressed as well when it comes to archiving data for later use, preferably something that is already done at the beginning of a research project (Humphrey et al., 2000). As an example, archiving data for video studies requires vast amounts of storage space and a well-organised data infrastructure. This is because data repositories from such studies often include digital files of student work, digitalized field notes, interview data, various metadata, and other digital resources, in addition to video data (Derry et al., 2010). Pea and Hay (2003) argue that developing effective metadata schemes is a central issue for the video research communities — if such data is to be exploited to their fullest potential. Associating some type of metadata to the video is a central step in the analysis of video data (Pea & Hay, 2003), and also in giving structure to a data repository. This is particularly important in archiving data for reuse, as it enables the secondary researchers to navigate and build on the archived data and metadata available to them.

Finally, it is important to consider carefully the access controls that are to be provided to secondary researchers. Access could be granted simply upon registration of a user, or there could be a requirement that permission must be requested. Access is also related to other activities that can be undertaken by secondary researchers, for example, tagging for datamining purposes, downloading, copying, etc. Decisions must be made about all of these possible functionalities and to whom they are permitted.

Cultural Challenges

Secondary analysis also requires also requires addressing cultural challenges in researcher attitudes. The act of data archiving requires substantial changes in research practice and beliefs about the nature of data. For example, many researchers view data collected from their projects as their private property. As a result, there is often very little to no culture of data sharing within a majority of disciplines or institutions. Researchers find it rare to envision how one's research could be based on data generated or created by another researcher. Data hoarding is often deemed to be a cultural norm. Some researchers fail to see any benefit in storing research data in a repository or providing secondary access to their data. Finally, researchers need to generate as much of their own scholarly output as possible from their dataset, and find it hard to let others benefit from the results of their labours (Bishop, 2009). Ultimately, numerous beliefs about secondary analysis and access to their 'private' data can be changed if researchers are reassured of the many benefits listed in the preceding section that can be accrued to them if they should decide to share their data.

Methodological and Epistemological Challenges

Methodological challenges to secondary analysis include assessing the project data for validity and errors before using them, considering the methodological rigour of the project, as well as

data limitations if any. For example, when considering a potential survey for secondary analysis, it is necessary to subject its methodology to critical scrutiny, including the quality of developmental and pilot work, interviewer training and fieldwork or online control, the method of sample selection, nature of the sampling frame and response rate. The secondary researcher needs to obtain as much documentation as possible about the collection of the survey data and be aware of any potential data limitations. An absolutely crucial task is to read and become familiar with the questionnaire and detail of wordings and response alternatives in the original survey. It is all too easy to begin to analyse variables in the dataset without really knowing exactly what the numbers represent.

Even more pertinent methodologically is with regards to qualitative data. The context and method of collection for qualitative data makes the archiving and reuse of such data methodologically more difficult: whereas quantitative data are obtained through abstraction from a context, qualitative data are highly contextualised (Gillies & Edwards, 2006; Kelder, 2006; Van Den Berg, 2006). For example, Mauthner et al. (1998) argue that the conditions and contexts under which data are produced are inescapable, rendering reuse of qualitative data problematic. Moore (2007), on the other hand, claims that that the labels of reuse and use create a false distinction between primary and secondary use of data, because all data are constituted, contextualised, and re-contextualised within any study or research process. Hammersley (2010, Section 4.9) contends that the "recontextualization argument" fails to acknowledge that data, in some sense, exist prior to the research process, as well as being constituted and constructed within any study:

> Data are, then, in an important sense given as well as constructed: they are not created out of nothing in the research process, nor should we construct whatever inferences we wish to on the basis of them. At the same time, it is important to recognise that they are also constructed or produced in the course of research, and to be aware of aspects of this process that could be relevant to what would and would not be legitimate inferences from them.

According to Hammersley, then, the methodological issue of context can arise in any research project, but the risk is greater when using secondary data. To address some of the contextual issues of reusing archived data, Bishop (2006, 2007) argues that it is necessary to consider the interactional, situational, and cultural or institutional levels of context that apply to the data. The interactional level of context refers to what the secondary researcher is likely to discover about the interaction or conversation in the data material, without having experienced the specific context it occurs in. The situational level refers to the setting, or the "context" as is traditionally known in qualitative studies. For instance, this includes the persons present, their relations, the physical setting, and so on. The third level of context concerns the institutional or cultural factors influencing the research project at the time of data collection. In an educational research setting, this may include the national curriculum at the time of the observations, the political situation, and leading reform initiatives. Considerations of these levels will help the secondary researcher to make better sense of the data, how to interpret it in light of such contextual understanding, and crucially, how they can address the key question of fit: how a research question is likely to be answered with the secondary data (Hammersley, 2010).

Epistemological challenges include issues around the notion that data are inseparable from the underlying theoretical assumptions, goals, agenda. To give a concrete example, one key function of survey research is theory testing. The primary researcher designs a research instrument and collects data in order to test some theoretical hypotheses. These are operationalised into empirical hypotheses, and questions are developed to validly measure the theoretical concepts that constitute the various elements of the researcher's theory. However, the secondary researcher has to work with someone else's survey questions and assess whether these questions adequately measure the concepts used in the theory they wish to test. It may not be possible to measure the key theoretical concepts because appropriate questions have not been asked in the survey or the questions may not be valid indicators of the relevant concepts. In this case, the secondary researcher can be pragmatic and evaluate how well available variables stand in for those that he or she would ideally like to have, and to write a justification of this in his or her results.

In terms of theory development, which tends to have a closer alignment with qualitative data, Bishop (2007) argues for the appropriateness of secondary analysis to be mediated in light of a particular research question or set of questions. From this perspective, certain types of studies and empirical questions may lend themselves to secondary analysis whereas others may not. For example, if the secondary data has been catalogued as having low instances of dialogic teaching, then research questions that aim to provide descriptive understandings of dialogic teaching may not be appropriate. Normative questions which aim to understand how teaching could be improved in more dialogical ways might find better resonance with a dataset that have low instances, but may have rich contextual information on teaching practices.

Ethical Challenges

Ethical challenges revolve around ensuring confidentiality and anonymity of respondents and researchers, ensuring informed consent for secondary analysis, and evaluating the risks of opening up data access to other researchers or users.

Certain forms of data, such as video recordings, are very sensitive to exposing the identities of respondents, participants and researchers (Corti, 2000). A common option to enable reuse and protect confidentiality is anonymization, usually by removing identifying information or camouflaging real names or images. The key issue here is to agree on an appropriate level of anonymization, so that the data are not distorted, or their potential for reuse reduced (Corti, Day & Backhouse, 2000). In the case of video data, however, anonymization is not so easily accomplished, nor is it always appropriate to do so if they are to be subjected to new analytic perspectives or procedures. For example, if the participants' faces need to be filtered out or masked in a video recording, the video data may lose most of its value for the secondary researcher. What is important is that in satisfying anonymity requirements, a balance must be struck between the removal of key identifying characteristics of research participants and retaining the integrity and quality of the secondary data. Derry et al. (2010) propose that confidentiality to the research participants can still be protected in several ways, even with the non-anonymous nature of video data. Filtering and masking the identities of the participants is a possibility, albeit an expensive one, which in turn could compromise the data. In the

videographic study by Lefstein and Snell (2013), the video recordings are processed using a masking filter which removes identifying characteristics without losing too much detail in the video data (see below). Such recordings can then be shared in a public domain without compromising confidentiality.

Figure 3: Masked Video Recording to Anonymise Participants (Source: Lefstein & Snell, 2013)

As informants usually consent to being part of a study under the promise of confidentiality with respect to the research project and its members, there is also the question of informed consent for the secondary researcher to consider (Heaton, 1998). For example, how was consent originally obtained? Corti et al. (2000) emphasise the importance of issues concerning informed consent being resolved prior to data acquisition, which implies that the ethical challenges of reusing data applies to primary researchers as much as it does to secondary researchers. Derry et al. (2010) conclude that these are important issues to address to enable sharing and reuse of video data, for example, by developing and sharing practices for obtaining informed consent that protect the research participants and support the future sharing of video data. Importantly, advice concerning informed consent from legal and institutional bodies suggests that informed consent must be ongoing along with a mandatory opt-out clause. That is, it must be possible for participants to withdraw from the study at any point where he or she no longer feels comfortable with it. This of course becomes exceedingly difficult when consent is being requested for possible secondary uses, which, by the nature of the case, are unknown to the researchers gathering the data. A number of options exist to address this issue. With participants who have access to the internet, one possibility is that they be shown or have access to examples of archived data (Luff et al., 2002), and allowed to see and possibly edit the data in which they are represented. It is also possible to use the functionalities of digital archives to enhance the on-going consent process and research participants' choice of opting out of the research. Research participants who are interested could follow their data through different uses that are made, and opt out at any point that they feel uncomfortable. For all this to be possible, research

participants should be registered participants on data archiving sites, if they are interested in being so. Responsibilities for data do not all lie with researchers; instead, by the same token, the very process of informed consent which is a form of respect for the autonomy of individuals puts a certain amount of responsibility on research participants as well. On-going risk evaluation must be made to ensure that risks to respondents, participants and researchers are minimised in the process of opening up access to data for secondary analysis.

Legal Challenges

There are also concrete concerns around intellectual property (Mauthner et al., 1998; Parry & Mauthner, 2004). Interviews, for example, are the intellectual property of both the interviewer and the interviewee, creating complex issues around allowing wider access to the data. Data can constitute a contract offering a contribution of intellectual property for pre-specified use by pre-specified users (for example, publication by the individual researcher or research team). This notion of ownership, where data are seen as produced by the researcher in the same way as outputs and hence subject to intellectual property regulations, has important implications for the sharing of data; it is likely that a standard blanket statement such as one that transfers ownership from researcher to the institute upon project closure may not be legally tenable. Additionally, the terms and conditions stipulated in the current OER secondary analysis application form are laden with legal inconsistencies that need to be subjected to proper legal deliberation. As a further example, it should be a legal condition that the following applies:

> Secondary analysts must be bound by the same confidentiality and privacy restrictions as the primary analysts and are to be held legally responsible for any breach of ethics in the conduct of secondary analysis.

To implement such a condition requires modifications in the wording of contracts as well as participant consent forms. Such modifications should clearly state that (a) data collected may be used for specific multiple research and training purposes, and if it is so used, (b) the secondary users are subject to the same legal and ethical obligations regarding confidentiality and privacy set forth for the original researchers.

Conclusion: Recommendations Moving Forward

Whilst good data management is fundamental for high quality research data and therefore research excellence, it is crucial for facilitating data sharing and ensuring the sustainability and accessibility of data in the long term and therefore their reuse for future research. If research data are well-organised, documented, preserved and accessible, and their accuracy and validity are controlled at all times, the result is high quality data, efficient research, findings based on solid evidence and the saving of time and resources.

Researchers themselves benefit greatly from good data management. Good data management should be planned before research starts and may not necessarily incur much additional time or costs if it is engrained in standard research practice. As it stands now, the responsibility for data management lies primarily with researchers, with many not practicing good data housekeeping or management, but institutions and organisations can provide a supporting framework of

guidance, tools and infrastructure and support staff can help with many facets of data management. Establishing the roles and responsibilities of all parties involved is key to successful data management and sharing. For example, Monash University's data management protocol is exemplary:

> At the design phase of a research project, researchers should undertake a process of decision-making and documentation of key research data management activities… While this is not yet common practice in Australia, it is likely in the near future that the Australian Research Council and the National Health and Medical Research Council will require greater evidence of data management planning. The Australian Code for Responsible Conduct of Research (2007) requires aspects of data management such as ownership, ethics, and retention and disposal to be well-documented by researchers… Other requirements—for example, secure storage and backup of digital data—are not currently well-documented and this can cause problems as personnel, technologies and research processes change over time. (http://www.researchdata.monash.edu/ guidelines/planning.html)

Accordingly, before a secondary research embarks on the road towards obtaining data, he or she should try to get answers to the following questions about the data (Stewart, 1984). Such questions can also help data archive administrators and project managers determine what sorts of information are required from the original researchers in order to make any archived data useful. The questions are:

- What was the purpose of the original study? Was it designed around a specific hypothesis which limits its usefulness for a researcher with different ideas?
- What information was collected? Were key variables defined in such a s way as to be compatible with the new analyst's ideas?
- What was the sample design and what sample was actually achieved? Is it adequately representative of those groups that play a key part in the new researcher's theory?
- What are the credentials of the data? Can the original researchers be assumed to be competent? What evidence is available about the reliability and validity of the data?
- When were they collected? Is topicality important or is it reasonable to assume that the relationship being studied are relatively invariant across time?
- Are the data nationally representative? Is this crucial, or is it reasonable to assume, as under the last point, that the relationships hold across a wide range of geographical localities?

It is important to consider, at the same time, what data should be archived for secondary access, as not all data may necessarily be useful or may have a strong fit with secondary research. The following guidelines, which focus on long term future value of research studies, can be used to determine the value of data for archiving and secondary access purposes:

- The historical value of the study.
- Data that are complementary to existing data holdings.

- Data that have further analytic potential than the original investigation, that is, they have not been exhaustively analysed.
- Data that are based on large-scale national and/or representative samples.
- Data which are longitudinal in design.
- Data which facilitate the possibility of further follow-up of the sample.
- Data that employ mixed methodology.
- Studies that include a wide range of measures.

Further guidelines that are established internationally, that can be used to develop data archiving processes, principles and resources, can be found in the Organisation for Economic Co-operation and Development (OECD) *Principles and Guidelines for Access to Research Data from Public Funding* (OECD, 2007) and the Inter-University Consortium for Political and Social Research (ICPSR) *Guide to Social Science Data Preparation and Archiving* (ICPSR, 2012). The OECD suggests that the principles of openness, flexibility, transparency, legal conformity, protection of intellectual property, formal responsibility, professionalism, interoperability, quality, security, efficiency, accountability and sustainability be upheld when designing systems for access to research data; many of these are discussed in the preceding sections. Ultimately, despite the potential pitfalls and challenges of secondary analysis, the international consensus is that secondary access to research data increases the returns from public investment, reinforces open scientific inquiry, encourages diversity of studies and opinion, promotes new areas of work and enables the exploration of topics not envisioned by the original researchers. These are benefits that can be readily accrued from the setting up of the Core 2 research data for proper secondary access, the key objective of the development grant.

Acknowledgements

This literature review was developed as part of the NIE funded project This study was funded by the Education Research Funding Programme, National Institute of Education (NIE), Nanyang Technological University, Singapore, (project no. DEV 01/14 RS). The views expressed in this paper are the authors' and do not necessarily represent the views of the NIE.

Appendix One: Summary Description of Core 2 Research Project

The Core 2 Research Project is a comprehensive, systemic baseline study of pedagogical and assessment practices in Singapore. It employs a large-scale, representative sampling that allows for generalizable evidence-based findings across the education system. The Core design utilises a coherent programme with nested structures linking different Panels.

The project has the following research questions:

1. How do teachers teach in Singapore?
2. What is the character and intellectual quality of instructional practice in Singapore?
3. Why do teachers teach the way they do?
4. To what extent has pedagogical practice changed since the introduction of *TLLM* in 2005?
5. What impact does instruction have on student achievement and the development of 21st Century skills?
6. What factors constrain the ability of the system to secure substantial and sustainable pedagogical improvements?
7. How might the quality of teaching and learning in Singapore be improved, given Core 2's research findings?

Research Design & Methods

The Core 2 project is organised around three separate but nested research projects or "panels": Panels 2, 3 and 5. Their design characteristics are summarized in Table 1 below.

	Panel 2	Panel 3	Panel 5
Design	Pre-test/Post-test non-experimental design	Cross sectional	Cross sectional
Sample	Schools: 62 Classes: 454 Students:16895 Teachers: 2100	Schools: 31 Lessons: 625 Units of Work: 117	Schools: 31 Lessons: 625 Units of Work: 117
Type	Survey of P5 and Sec 3 students. Survey of teachers	Classroom observation, coding and analysis	Collection and analysis of instructional tasks (n=385), student work (2,897), 115 teacher interviews, 209 T surveys
Assessment of student learning	Pre and Post assessments of all P5 and Sec 3 in Mathematics *or* English		Student homework, assessments, projects, class work
Data	Ordinal, interval	Binary / categorical; ordinal	Ordinal, interval
Unit of Analysis	L1: Individual students and teachers; L2: Classrooms and Schools	Teachers, units of work, lessons, lesson phases, exchanges/ interactions	Teachers, tasks and student work
Analytical Procedures	Descriptive statistics; exploratory and confirmatory factor analysis; correlational analysis; multiple regression; structural equation modelling; multi-level structural equation modelling	Descriptive stats; exploratory factor analysis; discourse analysis; conversation analysis; correspondence analysis; probability analysis (odds ratios); log linear analysis; Cox regressions; survival analysis; latent class transition models; network analysis; Markov Chains modelling	Descriptive statistics; exploratory factor analysis; regression analysis

Table 3: Research Design of Core 2 Project

Figure 1 below shows the research focus of each Panel of the Core 2 project:

Research Focus by Panel

	Panel 2 (Surveys)	Panel 3 (Clm. Obs)	Panel 5 (T&SW)
Students: A Descriptive and Multivariate Profile	x		
Teachers: A Descriptive and Multivariate Profile	x		x
Instructional Tasks	x	x	x
Task Fidelity (Task Implementation=Task Set Up)		x	x
Instructional Strategies (Generic)	x		
Instructional Strategies (High Leverage)	x	x	
Classroom Organization	x	x	
Structure of Classroom Interaction	x	x	
Classroom Knowledge Talk	x	x	
Instructional Profiles of Units of Work		x	
Case Studies of Exemplary Units of Work		x	
Modelling Impact of Instruction on Student Learning controlling for family background and student characteristics	x		
Modelling 21st Century Learning Resources	x		
Streaming, Instructional Practices and Student Outcomes	x		7

Figure 4: Research Focus of Individual Panels of Core 2 Project

Key Theoretical Perspectives

The Core 2 project draws upon a relatively large number of research areas and theoretical perspectives, as listed below:

- Pedagogical theory (descriptive, normative)
- Organizational theory, particularly new institutionalist theory
- The sociology of human capital formation and workplace organization
- Epistemology (normative epistemology, social epistemology, virtue epistemology)
- Disciplinarity Studies (intellectual history, history and philosophy of science, sociology of knowledge, functional systemic grammar)
- Learning sciences (cognition and learning theory, research on expertise)
- Research on motivation, engagement, metacognition, the self and self-beliefs (self-efficacy, self-concept, self-construals, self -regulation)

- Research on effective instruction, particularly research on instructional tasks, high leverage instructional strategies and classroom talk
- Research on mathematics education
- Research on language instruction
- Research on multi-modality
- Teacher effects research
- Educational production function analysis
- Statistical methodology, including confirmatory factor analysis, multi-level structural equation modelling, latent class transition modelling, network analysis, log-linear analysis
- Discourse and conversational analysis

Appendix Two: Possible Research Projects Utilising Secondary Analysis of the Core 2 Research Data

The following are examples of research projects that are currently utilising secondary analysis of the Core 2 research data. These are drawn from approved applications for access to the Core 2 data, with some applicants having prior experience with the Core 2 Panel 3 data. For these applicants, the issue of missing contextual information from the Core 2 data archive are not an issue as they were involved in the data collection and analysis stages of the Panel 3 research work. For other applicants, there is a need to provide detailed information and background on the Core 2 project to help them make sense of the data and to assist in their evaluation of the fit between their research questions and the secondary data.

Example 1: Student-initiated feedback practices in secondary 3 English Language classrooms in Singapore.

Research Questions:

1. What is the frequency; nature; and context of students' feedback practices in Secondary 3 EL classrooms in Singapore?
2. To what extent and how successfully can Hattie's model of feedback (2009) be applied to English language teaching and learning at the Secondary 3 level?
3. To what extent and in what way do specific intervention strategies impact students' feedback practices in Secondary 3 EL classrooms?

Example 2: Multimodality and meaning making in reading comprehension instruction in Singapore English secondary classrooms

Research Questions:

1. What kinds of classroom talk take place in meaning-making discourses of reading comprehension instruction?
2. What kinds of scaffolding do teachers provide when students are engaged in meaning-making discourses in reading comprehension instruction?
3. How does gesture contribute to the teacher's and students' ability to construct meaning from their engagement with reading comprehension texts?

Example 3: Dialogic teaching in Singapore English Language classrooms

Research Questions:

1. What does authentic dialogue (questioning) / dialogic instruction / teaching mean specifically for EL?
2. What does "dialogic" mean for an EL classroom?
3. What is the concept of dialogic interactions for EL, is it the same as for other subjects (Science, Math, Social Studies)?

4. How is meaning and understanding developed at EL lessons when language is both a tool and destination.
5. What kinds of understandings are developed in EL?
6. For an English Language unlike Science lesson mere introduction and explanation of an idea/concept by the Teacher is not enough. What are the points for understanding in EL (Because language is not something which can be learnt through transmission or just by introducing an idea, what kind of meanings do learners need to develop for EL and what does meaning making process for EL entail)?
7. What are the stages of meaning making for developing understanding and learning (on the inter-psychological plane of the classroom) for EL?
8. Is there a place for meaning-making and developing understanding through dialogue in EL if language is about rules and patterns? If yes, then how it is happening at EL class?
9. What is the tension/balance between Teacher authority and the possibility for a dialogue for students learning in EL?
10. How are classroom interventions carried out by a teacher in an EL class? What are the forms of pedagogical interventions specifically for EL? How (if at all) does the teacher intervene at EL lessons for meaning making and developing understanding in student learning?

Appendix Three: Sample Survey Items from Panel 2

The following are samples of the survey items from Panel 2 of the Core 2 Project.

Background information
Race (RACE)
Gender (GENDER)
Designation (DES)
Highest level of formal education completed (HEDU)
Name of school (SCHOOL)
Years of teaching experience (TEXP)
Year & Month of birth (BOY/BOM)
Teaching streams (STREAM)
Undergraduate major (UDMAJ)
Undergraduate **minor or second major** (UDMIN)
Post-graduate qualification (PGQE)
Teaching level and subjects (LVSUB)
level and Subject survey is based on (SLVSUB)
Teachers' Beliefs
Efficacy for Motivational Strategies (EMS)
Efficacy for Instructional Strategies (EIS)
Efficacy for Classroom Management (ECM)
Teachers' practices
Time distribution: Weekly (TDW)
Lesson preparation (PREP)
Communicating lesson goals, assessment standards (COM)
Learning Priorities (PRIOR)
Attribution Beliefs
Ability (ABIL)
Effort (EFFORT)
Learning and teaching
Conceptions of Teaching and Learning (TL)
Classroom Talk - Frequency (CTF)
Classroom Talk - Value (CTV)
Group Work (Group)
Differential instruction
Homework Feedback (FB)
High stakes summative assessment system (HSA)
Formal and informal assessments(IFA)
Professional Development and the School as a Professional Learning Community
Reflective dialogues (RD)
Professional collaboration (PC)
Quantum of Professional Development (PDQ)
Impact of Professional Development (Substantial and Sustained) (PDSS)
Influence on teaching and assessment practices (INFL)
School Culture, Organization and Leadership

Focus on student learning (FSL)
Providing intellectual stimulation (PIS)
Support for institutional change (SPC)
Shared decision making (SDM)

Table 4: Teachers' Survey for Panel 3

Background Information
Name of School [SCHOOL]
Class [CLASS]
Gender [GENDER]
Ethnicity [ETHN]
Stream [STREAM]
Birth year/month [BIRTH]
Parents Highest Education [PEDU]
Residence [RSD]
Father Language, Mother Language [FLAN] [MLAN]
English Listening, Speaking, Reading, Writing [ELS, ESP, ERD, EWR]
PSLE MATHS, PSLE English, PSLE Total [PSLEM, PSELE, PSLET]
Classroom Ranking/Position [RANK]
Self-Regulated Learning Beliefs
External Goals [EXTG]
Saliency of Parental Expectations [SPE]
Mastery Approach Goals [MAPG]
Mastery Avoidance Goals [MAVG]
Performance Approach Goals [PAPG]
Performance Avoidance Goals [PAVG]
Task value--Usefulness
Task value--Interest
Self-Efficacy [MEF]
Memory [MMR] + Reasoning [RSN]
Attribution [ATTR]
Math Anxiety [MANX] (PISA items)
MATHS Self-Concept [MSC] (PISA items)
Self-Regulated Learning Behaviours
Individual Engagement [IEA]
Group Engagement [GEA]
Time Management[TM]
Effort Regulation [EFW]
Homework Regulation [HWR]
Meta-Cognitive Self-Regulation [MCSR]
Surface Learning [SURF]
Deep Learning [DEEP]
Social-psychological outcomes
Affect at School [AFSH]

Self-Report of Disruptive Behaviour [DSBH]
Peer relationship [PERL]
Self-perception
Academic self-concept [ASC]
Educational Aspirations/Expectations [EASP]
Self-Determination [SDTM]
Interdependent Self-Construal (Harmony) [ITSC]
Independent Self-Construal (Uniqueness) [IDSC]
Contingency on Others' Approval [CGOT]
Agentic Self Concept [AGSC]
Some more 21st skills
Oral communication [ORAL]
Conscientiousness (reliability, punctuality, perseverance) [CNSC]
Problem Solving Self Efficacy Beliefs (PSEF)
Inventiveness (INVN)
Leadership skills [LEAD]
Teamwork skills [TEAM]
Coping/ Resiliency (COPE)
Family context
Family Resources [FR]
Family Learning Resource [FLR]
Family Support for Learning [FSPL]
Parental Warmth [PW]
Parental Structure [PS]
Parental autonomy [PA]

Table 5: Secondary 3 Student Survey Part 1

Background
School [SCHOOL]
Class [CLASS]
Gender [GENDER]
Ethnicity [ETHN]
Stream [STREAM]
Birth year/month [BIRTH]
PSLE English/Math/Total [PSLEM, PSLEE, PSLET]
Computer Access [CPAC]
Pedagogical Practice
Epistemic Pluralism [PLR]
Epistemic Norms [NORM]
Epistemic Virtues [VIRT]
Student Product [SPRD]
Goals (GOAL) +Criteria (CRTR)
Lesson Organization [LORG]
Structure and Clarity [SNC]

Flexibility [FLEX]
Information Communication Technology (ICT)
Collaboration [CLB]
Collective Feedback [FBCL]
Personal Feedback [FBPR]
Scaffolding [SCFD]
Understanding [UNSD]
Self-Directed Learning [SDL]
Self-Assessment [ASMS]
Peer-Assessment [ASMP]
Questioning [QU]
Curiosity and Interest [CNI]
Classroom Task Goals [CTG]
Classroom Ability Goals [CAG]
Positive Classroom Climate [PCC]
Authentic Pedagogy [AP]
Traditional Pedagogy [TP]
Good Teacher [GT]
Behaviour Management [BM]
Maximize Learning Time [MLT]
Review [REV]
Pace [PACE]
Quality of Homework [QHW]
Teacher's Support [TS]
Peer Academic Orientation [PACD]
Math Pedagogy [MP]
Exam Preparation [EXPR]

Table 6: Secondary 3 Student Survey Part 2