

**From assessment to implementation: Creating a standardized data competency measure and discipline-based RDM module**

**Abstract**

Texas A&M University (TAMU), in partnership with the University of Oklahoma (OU), Purdue University's Libraries, and University Libraries of the University of Nevada, Reno, Texas State University (TXST) Libraries, and Prairie View A&M University Libraries (PVAMU) is seeking an IMLS National Leadership Grants for Applied Research project. The proposed \$465,648 applied research grant, built upon two internal grants from TAMU, has the potential to bring about a significant impact on the research community and empower librarian professionals.

The grant will be used to **design a standardized measure/survey of data core competencies for social science, with a particular focus on education and library and information science (LIS)**, which will be created through a combination of environment scan data and student survey, as well as faculty interview and student focus group data. Additionally, the project will promote the sharing of knowledge and best practices in the field of research data management through the development of **an open access, discipline-based, and evidence-based Research Data Management (RDM) module**, which can be utilized in social science research method courses, as well as research data management service at academic libraries and RDM courses at LIS schools.

This applied research project will answer the following research questions: **1) What are the core components for the standardized data core competencies measure/survey for social science? 2) What is the status of social science graduate students' data core competency? 3) What is the effect of the discipline-based RDM module on social science graduate students' RDM skills? 2) Does the effect of the discipline-based RDM module vary by students' department, minority status, and years in the program?**

This project aligns with National Leadership Grants for Libraries Program Goal 2 and Objective 2.1 and 2.3, and supports the overall mission of IMLS to advance, support, and empower America's museums, libraries, and related organizations. The impact of this project includes the creation of a widely applicable standardized measure/survey for data core competencies that will have far-reaching effects on the social science and LIS research communities, as well as on the quality of academic libraries' RDM services in supporting student success. The project will run from August 1, 2024 - July 31, 2027.

**1. Project Justification**

**1.1. Statement of National Need**

Over the past decade, the importance of RDM has gained widespread recognition for its significant contributions to academia, government, and industry <sup>[1-3]</sup>. With the increasing use of data-driven methods across various fields of research, there is a growing demand for better access and organization of digital data. This demand is expected to rise as the volume of available data increases. According to the new guidance of the Office of Science and Technology Policy (2022) <sup>[4]</sup>, there is an immediate need for public access of scientific data to federally funded research, and researchers should be guided on their responsibilities regarding data management and sharing plans. Therefore, effective RDM is crucial to support the research process and ensure the integrity, accessibility, and reuse of research data.

Despite the clear need for robust RDM practices, there is a notable educational gap, particularly in non-STEM disciplines, where RDM instruction is less commonly taught. Shao et al.'s thematic analysis of curricular mapping in data-related courses illustrates a strong emphasis on statistics, computer science, and domain knowledge education, but low coverage of ethics, data management, and communication & visualization <sup>[5]</sup>. Klenke et al. found that RDM is the least commonly taught in non-STEM disciplines than in STEM disciplines <sup>[6]</sup>. These findings were supported by examination of

six social science disciplines such as Criminal Justice, Geography, Geology, Journalism, Political Science, and Sociology. Additionally, graduate students have also been identified as in need of ongoing data management instruction<sup>[7, 8]</sup>, but current RDM instruction for this group is still in its early stage. Few studies have reported specifically tailored RDM instruction for graduate students in social science disciplines<sup>[1, 9]</sup>. The lack of RDM education at the graduate level in social science disciplines is a persistent issue that this proposal aims to address.

The disciplines of education and LIS play a pivotal role in this context. Education is central to the development of individuals and societies, exploring the processes by which knowledge and values are transmitted and their impact on societal structures, identities, and power dynamics<sup>[10]</sup>. LIS, similarly, is crucial for organizing and disseminating information, making it accessible and understandable to the public. This discipline's focus extends beyond mere information storage and retrieval, encompassing the critical evaluation of information sources, information literacy, digital literacy, and the ethical use of information<sup>[11]</sup>. The proposed project leverages the synergy between LIS and education within social sciences and the expertise and experience of its team members in these areas to address the national need for enhanced RDM practices.

To effectively train graduate students in RDM, it is of critical importance to first assess their data core competency status. While several studies have attempted to identify the necessary skill sets for various disciplines<sup>[12-14]</sup>, the lack of standardized measure/survey of data core competencies in the social sciences poses a significant challenge (**Detailed information can be found in 1.2**). Addressing this, there is an imperative to construct reliable and valid measures specifically designed for the nuanced demands of social science graduate students. Establishing these benchmarks will not only facilitate a precise assessment of their competency levels but also enable the creation of tailored RDM training that directly targets their unique needs and fills existing skill voids.

This endeavor not only responds to a pressing national requirement but also sets a precedent for interdisciplinary collaboration that enhances the capacity for managing the vast amounts of data generated by social science research. This initiative stands as a testament to the importance of integrating LIS principles with educational research to advance the collective understanding within these fields, positioning the proposal as a pioneering effort in social science scholarship and addressing current policy imperatives for data transparency and accountability.

## **1.2. Previous Research and Practice**

Over the last two decades, there has been a growing demand for RDM services in academic libraries<sup>[15]</sup>, which include RDM training for librarians, faculty, and researchers<sup>[16-19]</sup>. Competency in RDM is necessary for all academic fields, and past research has shown that consistent data management training is beneficial for graduate students<sup>[7, 8, 18]</sup>. To provide effective RDM training, it is critical to understand students' experiences with and perceptions of RDM.

Data core competency refers to a broad set of skills required to effectively access, evaluate, transform, summarize, preserve, and present data. It is becoming increasingly important for researchers to possess these skills, as evidenced by reports indicating a growing need for managing research data. This includes understanding the infrastructure necessary for data sharing and reuse, as well as the social and cultural environments surrounding data sharing. To bridge the gap between experienced and early-stage researchers, there is an urgent need for university-level RDM education<sup>[20]</sup>. To achieve this, many researchers and librarians are calling for a graduate-level RDM curriculum<sup>[1, 9, 21]</sup>. **A standardized data core competency measure/survey is critical in the RDM field to design an effective RDM curriculum and to evaluate its effectiveness.** Several studies have already been conducted to survey and identify the data core competencies required by researchers and students<sup>[12, 14]</sup>. These competencies are vital to ensure that researchers have the necessary skills to handle data effectively in their work.

**The lack of standardized measure/survey of data core competencies in the social sciences is a significant challenge.** This is due to the notable differences in RDM practices between different disciplines, as highlighted in research by Cabrera et al. [22]. Although several studies have explored data core competencies in STEM fields, such research is limited in the social sciences and humanities. While disciplines such as Genetics and Physics have well-established data management practices and infrastructures [23-25], the humanities and social sciences are still in the process of developing data models, practice, and cultures around RDM (See **Supporting Document 1: Data core competency studies for different disciplines**).

This is particularly challenging in social science, where researchers handle a wide array of data types<sup>[26]</sup>, from large-scale digital repositories in LIS featuring diverse multimedia content to complex qualitative and quantitative data in educational research. LIS researchers face challenges in metadata organization, preservation, and access, while educational research delves into assessment data, standardized test scores, and quantitative data from interviews, large-scale survey data, auto-generated process data, experimental and behavioral data, audio-visual recordings, texts from official documents, and more<sup>[27]</sup>. Such work demands robust research methods and interdisciplinary collaboration to ensure data quality, interoperability, and long-term preservation. Additionally, social science data often relates to the disclosure of confidential or sensitive information about individuals or organizations. This complexity of social sciences data, and the diverse types of data involved, make it difficult to establish standardized data core competency measures specifically for social science.

It is of utmost importance to assess the data core competencies of social science researchers to evaluate whether they have developed the necessary skills to meet funding agencies' standards for data management and sharing. To do this, a standardized data core competency measure/survey should be developed to serve as an indicator of the researcher's actual level of data core competency. This measure/survey can be used to identify gaps in knowledge and skills, diagnose learning needs, and provide RDM instruction and support to all researchers. Additionally, the measure's implementation can lead to greater public awareness and policy changes regarding data management practices.

Despite the availability of open-access RDM sources the Data Management Training (DMT) Clearinghouse, DataONE Education, MANTRA, and a Coursera course that provides an introduction to research data management and sharing, these are generally not customized for the unique needs of disciplines such as Education and LIS. Moreover, while there are discipline-tailored RDM instructions available, most of them are concentrated on STEM disciplines [28-33]. The New England Collaborative Data Management Curriculum (NECDMC) is designed for undergraduates, graduate students, and researchers<sup>[34]</sup> but is also used primarily in the STEM disciplines. Only a few studies have focused on RDM instruction in the social sciences<sup>[1, 35]</sup>. However, given that social science fields significantly rely on secondary data, data from human beings, and both quantitative and qualitative research data<sup>[36]</sup>, it is crucial to create and provide RDM instructional modules tailored explicitly for social science subject areas. Additionally, it is essential to apply the standardized data core competencies measure/survey to evaluate the effectiveness of such RDM modules for social sciences.

### **1.3. The Solution**

Significant research gaps exist regarding the RDM instruction needs from the perspectives of the students – our next-generation researchers, especially for those in the social sciences. Addressing this gap, our project proposes two key initiatives. First, this project aims to develop a standardized data core competency measure/survey for social science, focusing on education and LIS. This tool will enable librarians to evaluate and improve their own RDM services and curricula, ensuring that they are providing high-quality RDM services that are aligned with the latest best practices and standards. Second, the proposed study will generate a discipline-based and evidence-based open access RDM module. This module will be a valuable resource for library RDM professionals and information science faculties, as well as for the social science research community. This effort seeks to make a

significant societal impact by establishing a universal data core competency framework and fostering RDM knowledge sharing, aligning with the goals (2) and objectives (2.1 & 2.3) of the NLGLP and supports the mission of IMLS to support and empower American museums, libraries, and related organizations.

Our initiative is poised to be a pioneering effort in introducing a standardized measure/survey of data core competency in social science. We intend to rigorously develop and validate this measure/survey through empirical research and experimental design, focusing on enhancing RDM education for graduate students in these fields. The development will be theory-driven and involve a panel of experts, as well as rigorous psychometric evaluations using scientific analyses such as exploratory analysis (EFA). Additionally, we plan to conduct experimental studies with RDM modules into graduate-level research methods courses. Through this project, we hope to contribute to the development of effective RDM education for social science (focusing on education and LIS) researchers and provide library professionals a useful tool to support students' success, filling an important gap in the field.

## 2. Project Work Plan

### 2.1. Project Design

This three-year project will use a mixed-method research approach to design a standardized measure/survey for data core competency, design a discipline-based and evidence-based RDM module, and conduct an experimental/intervention study via pre- and post-design in graduate students' research methods course.

### 2.2. Research Questions

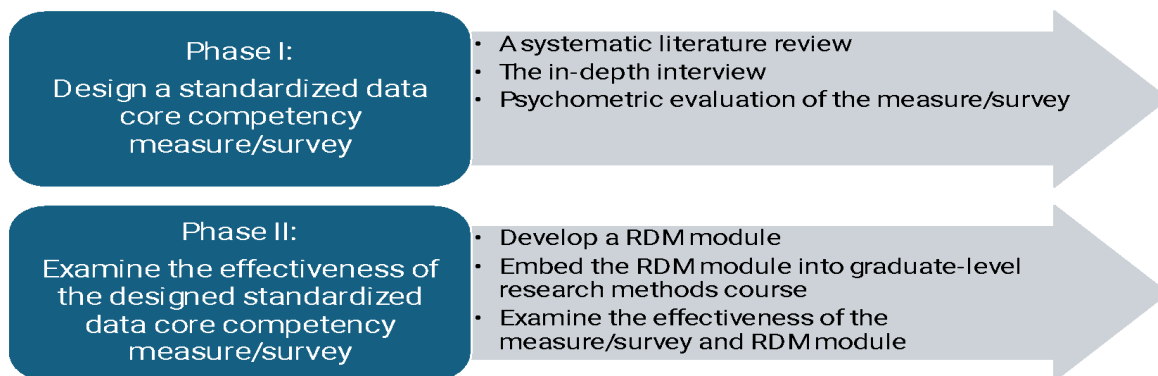
This research project will investigate the following research questions:

- 1) *What constitutes the essential components of a standardized data core competencies measure/survey for social science?*
- 2) *What is the status of social science graduate students' data core competency?*
- 3) *To what extent does the designed standardized measure/survey for data core competencies, as evaluated through an experimental study (a discipline-based RDM module intervention on social science (library science and education) graduate students), capture the changes in students' data core competencies?*
- 4) *Does the impact of the discipline-based RDM module vary by students' department, minority status, and years in the program?*

### 2.3. Project Work Plan

This project can be divided into two major phases (**Figure 1**):

**Figure 1. Project work plan.**



Phase 1 – **Design a standardized measure/survey of data core competencies (Aug. 2024 –May. 2025)**

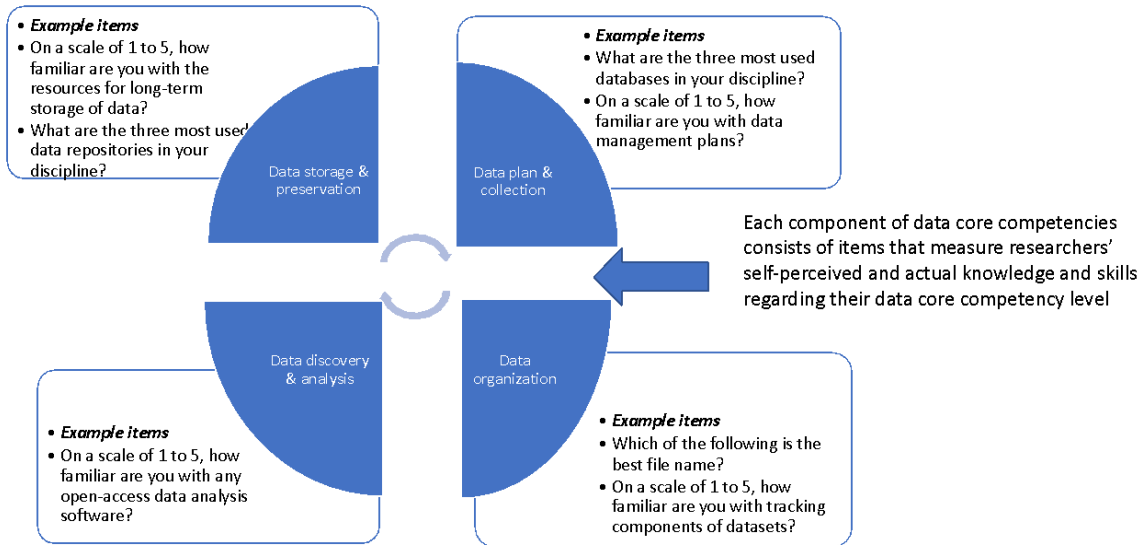
- A **Systematic Literature Review (Aug. 2024 - Aug. 2025)** will be conducted to gather a comprehensive item pool of data core competencies from the relevant literature. Dr. Xu’s team from TAMU, Dr. Zhou from TXST, Dr. Zakharov from Purdue, and Dr. Abbas from OU will utilize a scoping study framework <sup>[1, 37]</sup> to review peer reviewed studies and conference proceedings from Jan. 2010 to Aug. 2024. The purpose for this systematic literature review is to **identify the relevant literature concerning data core competency and corresponding assessment for data core competency**. The systematic search of literature will be retrieved from Library, Information Science and Technology Abstracts ([LISTA], EBSCO), ERIC, and Medline Complete, as well as conference proceedings (i.e., ACRL, ASEE) and others key journals in the field that have not been indexed by the databases. This will enable us to compile a comprehensive item pool of data core competencies from literature, building upon previous research outcomes.

While conducting the systematic literature review, an Institutional Review Board (IRB) submission, consisting of the recruitment script, consent forms, and all research instruments (survey, interview questions, and focus group questions), will be submitted for the TAMU IRB for approval.

- **The in-depth interview phase (Dec. 2024 - Aug. 2025)** will involve conducting interviews with 10 experienced faculty members and four focus group discussions with 20 to 30 graduate students in the library science and education fields. Our objective is to identify the components of data core competency for social science from their perspectives. We will involve researchers from six collaborating institutions, including both tier 1 university and tier 2 universities, aiming for a comprehensive understanding from both seasoned and emerging researchers. Our selection criteria will emphasize diversity in career stages, demographic backgrounds, and a commitment to including at least 20% of underrepresented minority members. This approach ensures a varied collection of insights across different disciplines, methodologies, and data types within LIS and education. **The semi-structured and in-depth format of the interviews and focus group is strategically designed to capture the nuanced views of established researchers on data core competency, while also identifying the RDM needs of emerging researchers.**

Dr. Xu (TAMU), Dr. Zhou (TXST), Dr. Zakharov (Purdue), and Dr. Abbas (OU) will design the first draft of the standardized data core competencies measure/survey. This initial design will be informed by a systematic literature review, alongside interviews with social science faculty and graduate students. The measure/survey will cover essential areas such as data plan and collection, data organization, data discovery and analysis, and data storage and preservation. It will assess both the self-perceived and actual knowledge and skills of researchers in these core data management areas. The draft will then be refined based on feedback from Ms. Schultz (UNR), Dr. Brumfield (PVAMU), and insights from the project’s advisory board, ensuring a comprehensive and effective tool for evaluating data core competency (**More detailed information can be found in Figure 2**).

**Figure 2. Conceptual design of the standardized data core competency measure/survey.**



*Note: This represents the conceptual design for standardized data core competency measures. However, the number of components included may be subject to adjustment based on the findings from the literature review and interviews conducted.*

● **Psychometric evaluation (Sep. 2025 - May. 2026):** Using the first draft of data core competencies measure/survey, we will conduct the survey with a stratified sample ( $N > 300$ ) from across six collaborating institutions using Qualtrics. To ensure diversity in the sample, each institution will distribute the survey with the same discipline during the same period. **The stratified sampling strategy will be used to maintain diversity in the participants and promote representation of underrepresented minority students** [38, 39]. The target response rate is 10%, and a random sample of 400 students will receive a \$20 e-gift card as an incentive to participate.

To maintain consistency in the recruitment and data collection process across the six collaborating institutions and to ensure the survey participants are representative of the student body (such as their disciplines, program, and race/ethnicity), the project director (PD) will provide detailed written instructions on the recruitment plan and data collection strategies. Regular project meetings with the co-PIs from each institution will be held each month before distributing the surveys.

To improve the quality of our item pool and select the most suitable items for each data core competency, we will analyze respondent data using several techniques. First, we will remove any outliers from the item pool that do not accurately reflect the target population's true competency. We will then conduct a **parallel analysis** [40] to determine the number of key components required to reflect the essential aspects of data core competency. We will prioritize the results suggested by empirical data rather than relying solely on the literature review and interviews. Next, we will use **exploratory factor analysis (EFA)** to cluster the measured items and identify the items that are strongly related to each component of data core competency. Based on the EFA statistics, we will retain the items that have the strongest correlation with each competency component. We will also **revise items** if any issues are identified (e.g., wording or response options).

The data core competencies measure/survey will be improved for reliability and validity based on the analysis results. The TAMU team will lead the data analysis, with input and feedback from the project team and advisory board members. Our goal is to create a high-quality standard data core competency measure/survey for social science that accurately reflects researchers' knowledge and self-perceived abilities in data management.

**Phase 2 – Examine the effectiveness of the designed standardized measure/survey of data core competencies (Jan. 2026 – July. 2027)**

In the second stage of development of the standard measure/survey of data core competencies for social science, we will conduct a pre- and post-test design experiment to assess the effectiveness of the proposed standardized measure/survey.

- **Develop a RDM module (Jan. 2026 - Aug. 2026):** The project team, leveraging their extensive instructional expertise in RDM education (**Detailed information can be found in 2.4**), will develop a discipline-specific and evidence-based RDM module in Learning Management Systems (LMSs), using the findings from the psychometric evaluation. Additionally, the project team plans to leverage existing RDM educational resources<sup>[30, 34]</sup> to design a comprehensive RDM module. This effort will utilize materials from established platforms such as the Data Management Training (DMT) Clearinghouse, DataONE Education, MANTRA, and a Coursera course on RDM basics. By integrating these resources with the specific data core competency levels identified for graduate students, the module will offer tailored examples and content relevant to the targeted disciplines—Education and LIS. It will address key areas such as the management of qualitative and mixed data types, and the ethical considerations involved in handling data related to human subjects.

Expected to comprise 5~6 sections, the module will cover data planning and collection, data organization, data discovery and analysis, data storage and preservation, and data sharing. Each section will include a comprehensive set of materials, such as slides, lesson plans, video lectures, in-class exercise and after-class exercise, and teaching guides, to facilitate effective learning. To ensure the module's alignment with current needs and standards, the project team will engage in consultations with all team members and an advisory board. This collaborative approach aims to refine and enhance the RDM module, making it a valuable resource for fostering data management skills among graduate students in the fields of LIS and education.

- **Embed the RDM module into the core curriculum of graduate-level research methods course (Sep. 2026 - July. 2027).** The project team is actively collaborating with faculty members from library science and education to seamlessly incorporate the designed RDM module into the core curriculum of graduate-level research methods courses. This initiative is underpinned by substantial faculty support, as evidenced by the detailed endorsements outlined in **Supporting Document 2: Supporting letters from faculties**, which contains letters from education and LIS faculty members expressing their strong interest in and support for this project. To date, the project has received **17 supporting letters across 13 institutions from the field of Education and LIS**. These letters are a testament to the widespread faculty backing for this initiative, indicating a broad-based recognition of its potential impact on enhancing RDM skills for graduate students. We are confident that the support highlighted represents just the beginning of a growing consensus on the value of integrating RDM into academic curricula.

- **Examine the effectiveness of the designed standardized measure/survey of data core competencies and the RDM module (Sep. 2026 - July. 2027)** via a pre- and post-test design, using the measure/survey developed in Phase 1. Dr. Xu's TAMU team will work with Dr. Zakharov from Purdue and Dr. Zhou from TXST to carry out the experimental study, with feedback from the other team members. Statistical techniques such as one-way ANOVA, regression analysis, and T-test will be utilized to determine the impact of the intervention and answer research questions. The project team will collaborate with the advisory board to **finalize the RDM module to make it openly accessible.** (See **Supporting Document 3: Research questions and proposed research method and data analysis plan**)

## 2.4. Key Personnel

**Dr. Zhihong Xu (Principal Investigator and Project Director)**, is an associate professor in the Department of Agricultural Leadership, Education, and Communications at TAMU. Dr. Xu worked as a data management librarian at the University Libraries from 2020. Since then, RDM has been one of her core research agenda. With expertise in research data management and educational technology,

Dr. Xu is a recognized expert in her field, having published numerous peer-reviewed papers in high-impact journals. Dr. Xu teaches comprehensive RDM/data analysis courses and workshops at TAMU. Her role encompasses overseeing every aspect of the project, including survey and interview questions creation, data collection and analysis, mentoring graduate students, and dissemination of results. Dr. Xu's leadership is pivotal to the project's success, leveraging her extensive research and mentorship experience.

**Dr. Wei Zakharov (Co-Principal Investigator)** is an associate professor at Purdue University Libraries. Dr. Zakharov offers credit courses *Introduction to Data Management* and *Understanding Your Research Data* at Purdue. Her area of expertise lies in data and information literacy education and online learning. Her contributions will be integral to the design of the survey and interview questions, the RDM module, and data collection.

**Dr. June Abbas (Co-Principal Investigator)**, is a Professor and Director of the School of Library and Information Studies, and Co-Director of the Data Scholarship Program at the University of Oklahoma. Dr. Abbas teaches credit courses *Data Stewardship* and *Organization of Information* which includes elements of data management at OU. Her expertise is in knowledge organization, user-centered design, data stewardship, and survey design. Her responsibilities include survey design, data collection, and acting as a liaison with educational partners, playing a key role in the project's execution and outreach.

**Dr. Xuan Zhou (Co-Principal Investigator)**, a Data Curation Specialist at TXST Libraries. Dr. Zhou teaches Data Management 101 at TXST focusing on fundamental concepts and competencies in RDM. Dr. Zhou brings a unique perspective with her background in research data management and teacher professional development. Her involvement will enhance the project's methodological design and analysis, contributing to its overall effectiveness.

**Dr. Elizabeth Brumfield (Co-Principal Investigator)**, holds the position of PVAMU Distance Services Librarian and the Head of the Northwest Houston Center Library. Dr. Brumfield served on the Black Caucus of the American Library Association and the ALA Office of Diversity, Literacy, and Outreach Advisory Committee. Her insights will be vital in shaping the study's approach and ensuring its relevance across diverse groups.

**Teresa Schultz (Co-Principal Investigator)** is an associate professor at the libraries of UNR. Prof. Schultz contributes expertise in data literacy and RDM through various workshops and classes, emphasizing data understanding, finding, managing and sharing best practices. Ms. Schultz's contribution will be critical in ensuring that the study is well-rounded and addresses the data literacy needs of graduate students in a comprehensive manner.

The project is further supported by a highly qualified **Advisory Board** consisting of experts in their respective fields. The members of the board include **Dr. Oi-Man Kwok** from TAMU, a quantitative method and psychometric measurement expert; **Ms. Sonia Barbosa** from Harvard University, a data curation expert; **Ms. Abigail Goben** from University of Illinois Chicago, a data management expert; **Mr. Jason Clark** from Montana State University, another data management expert; and **Dr. Matt Baker** from TAMU, a program assessment expert. The advisory board members have committed to regularly reviewing the project and providing valuable feedback to the project team. Starting from December 2024, they will review the progress of the project twice a year, up until July 2027. This collaborative effort underscores a commitment to advancing RDM education and practices through interdisciplinary expertise and strategic partnerships (**See Supporting Document 4: Letters of commitment from Advisory Board members** for a more in-depth discussion of their interest and support for the project).

## **2.5. Performance Measurement Plan.**

The research project maintains the highest standards of research integrity, quality, and reliability throughout every stage. To achieve this, a comprehensive set of strategies and policies will



be put in place to ensure the accuracy and reliability of the research. These strategies include: establishing clear protocols for data normalization to ensure consistent data formats and measurement standards; implementing rigorous data handling and analysis procedures to guarantee the accuracy and reliability of the results; selecting appropriate data collection and storage tools that prompt data consistency and ease of access; adhering to best practices in metadata documentation to preserve the context and history of the data; providing comprehensive training for research staff to ensure they are equipped with the necessary skills to carry out their tasks effectively, etc. (**See Performance measurement plan table**).

To ensure the privacy and confidentiality of research participants, all data intended for sharing will be carefully deidentified. **The deidentified data**, including the data from five collaborating institutions, the standardized data core competency measure/survey for social science, the assessment tool, and the RDM open-access module, will be **securely deposited in the Texas Data Repository**. This deposit will ensure that the data is safely preserved and accessible to the research community for future studies. The project team takes data security and privacy seriously, and all necessary measures will be taken to protect the sensitive information of the research participants.

## **2.6. Communication and Dissemination Plan.**

The project team has planned a comprehensive dissemination strategy to share their research products and findings widely. This will include participation in at least four national and international conferences (e.g., ACRL, RDAP, SLA, & AERA). In addition, the team will publish at least five articles in leading open-access research journals (e.g., LISR, JASIST, PLOS ONE, C&RL). We will also use social media and other resources (e.g., OSF, TDR) to reach a broader audience. The reach of this project is immense, and it has potential to shape the future of RDM by influencing the next generation of researchers and promoting best practices in data management. To further promote our work, the team will host a series of webinars disseminating the standardized data core competency measure/survey and strategies for its use. These webinars will be promoted through library association platforms such as ACRL Online Discussion Forums, ACRL Presents, RDAP listserv, and SLA listserv. To make the standardized data core competency measure/survey more accessible to the public, the team will make it available through the Texas Data Repository. We will also develop an open-access RDM module, based on discipline and evidence, that will be made available to the public.

## **3. Diversity Plan**

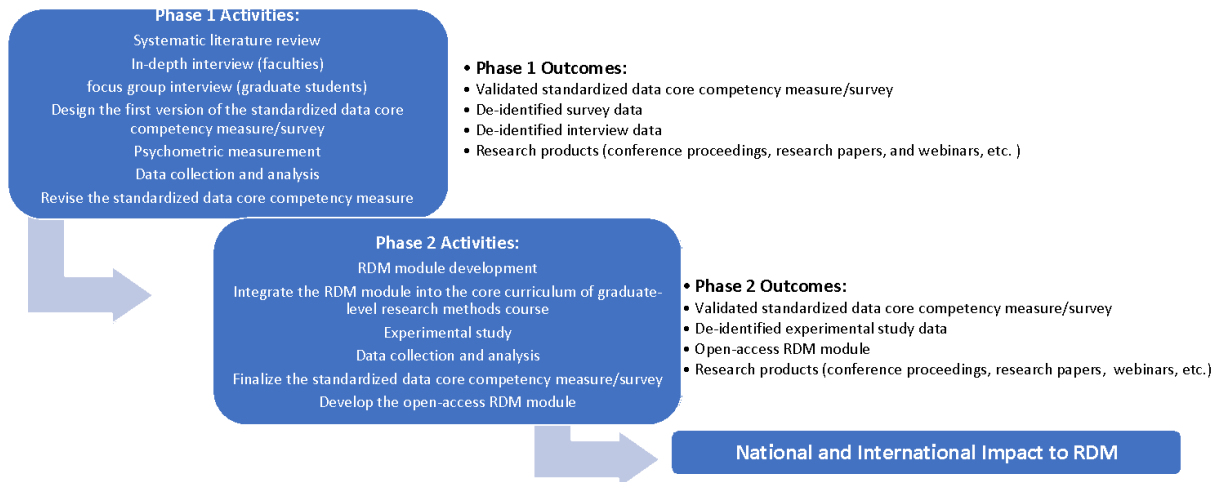
This project prioritizes diversity in developing an inclusive standardized data core competency measure/survey and RDM module through several key strategies: 1) Collaboration with Hispanic Serving Institutions (HSIs) such as TAMU, UNR, and TXST, as well as Historically Black Colleges and Universities (HBCUs) like PVAMU, ensuring diverse perspectives are represented in the project's development; 2) Enhanced representation in data collection, utilizing inclusive language and actively seeking feedback from underrepresented communities to ensure data is disaggregated by race, ethnicity, gender, and other relevant variables; 3) Incorporation of diversity in analysis and interpretation, considering how various demographic factors may influence data core competencies and addressing any disparities uncovered; 4) Ongoing solicitation of diverse feedback from stakeholders, including the advisory board, scholarly literature, and underrepresented communities, to continually enhance the project's equity and inclusion efforts.

## **4. Project Results**

### **4.1. Results.**

Outcomes of this project will be a standardized data core competency measure/survey for social science, conference proceedings and publications, publicly available revised and validated assessment tools, de-identified data shared for re-use, and the open access RDM module.

**Figure 3. Summary of the two-phase project activities and outcomes**



#### 4.2. National/International Impact.

Our project will make a groundbreaking contribution to the field of Research Data Management (RDM) by developing the first standardized data core competency measure/survey for social science. This measure/survey will be pivotal in identifying disparities in data competency, which is crucial in comprehending the varied data practices across different disciplines [35, 41]. The development of this measure/survey will lay a robust foundation for other disciplines, significantly expanding our understanding of researchers' data competency status across fields. This thoroughly tested instrument will be also able to allow the project team to potentially carry out a longitudinal evaluation of students' gains in data competency. Furthermore, if the instrument serves as a common program assessment tool across libraries' programs, essential data could be provided to IMLS about the national impact of research libraries on student learning and achievement.

The project's impact will be far-reaching, as the results will provide academic libraries with valuable insights into the RDM resources and services that are essential for student success. With this outcome, we will design a discipline-specific and evidence-based RDM module tailored for social science (focusing on education and LIS), which can be seamlessly integrated into learning management systems (LMSs). This module will greatly enhance the RDM knowledge of next-generation researchers, leading to a marked improvement in the quality of their research. Furthermore, this curriculum will be made available via Creative Commons licenses, and its effectiveness will be examined through its integration into the core curriculum of graduate-level research methods courses at partner universities, and more broadly, expanding the benefits to both higher education and library communities.

#### 4.3. Sustainability

This multi-institutional, multidisciplinary, and collaborative study project personnel diverse backgrounds and expertise, and expands readers' communities. The strategic partnership among six institutions, including four ARL members, and close affiliation with RDAPs working groups ensure widespread dissemination and adoption of project outcomes for interdisciplinary impact. To ensure long-term accessibility and preservation, data and products will be stored in standard file formats. Comprehensive documentation will support the long-term use and application of the standardized data core competency measure/survey developed in this project (**Detailed information can be found in the Performance measurement plan table and Data management plan**). The project team is committed to providing access to these resources for a minimum of five years after the grant period, with the expectation of indefinite access in the future (**See Supporting document 5: References**).

**Schedule of Completion - Phase 1 Design a standardized measure/survey of data core competencies**

Task	Specific activity	2024					2025												2026					
		Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	
Systematic literature review	Literature Search	x	x	x	x																			
	Coding & data analysis					x	x	x	x															
	Results report								x	x	x	x	x											
IRB application		x	x	x	x	x																		
In-depth interview	Recruit participants					x																		
	Conduct the interview						x	x																
	Conduct the focus group interview								x	x														
	Data analysis								x	x	x	x												
	Results report												x	x										
Data core competency measure/survey development	Item pool creation						x	x	x	x	x	x	x											
	Data core competency measure/survey version 1													x										
Psychometric evaluation	Recruit participants													x	x	x								



## Digital Products Plan

- **Digital Products Type**

Project resources will include an open access research data management module, environment scan results, survey data, interview data, data from the experimental study, conference presentations, and peer-reviewed publications. Survey data will be collected using Texas A&M University (TAMU) license of Qualtrics, then exported in .csv format. Interview data will be saved as .mp3 file and transcribed and saved in a .txt file. Experimental studies' data will be stored in .csv format as well. The other resources and data will be distributed and made freely available online in PDF format or MPEG-4 (.mp4) format. Standard web services will be used to create project resources, such as Google Docs. All project documents will be saved and shared using non-proprietary, openly documented file formats.

- **Availability**

To promote long-term preservation and open access to the data generated in the project, the project will use open formats to store all the documents. The use of open source software such as R and RStudio and Python, will be prioritized throughout the project, and R.script will be generated and stored for all data processing and analysis. Additionally, the project will create digital metadata, including readme files and codebooks, for each type of primary data.

The published documentation will include the information necessary to read and interpret the data, such as file structures, and instructions for R.script. Raw data will be preserved in a way that includes self-identifying information, which helps preserve provenance. Additionally, the project will use file naming conventions that provide a second level of visibility into the data's provenance. To ensure that the data is fully described and understood, the project team will manually generate descriptions and other metadata as appropriate.

- **Access**

All project data will be shared via the Texas Data Repository (TDR). TDR provides long-term preservation of digital objects using an off-site backup and assigns a Digital Object Identifier (DOI) for citations. The repository provides bit-level preservation and ensures ongoing access to research data, including associated metadata and documentation for a minimum period of ten years after it is deposited. All data will be retained for a minimum of three years after the conclusion of the project or public release (publication), whichever is later. Data related to a student's research work will be retained at least three years after the student has graduated. Access to the database and associated software tools generated under the project will be available

for educational and research purposes. Intellectual property rights will be retained by the universities and investigators.

To protect the privacy of participants in the survey and interview process, the project will implement a de-identification process for personal information obtained in the data collection. Firstly, the project will restrict access to the raw data and only allow PI or Co-PI to access sensitive information. Next, the project will utilize pseudonymization to remove identifying information from raw data. This will involve replacing identifiable information, such as names, age, and gender, with randomly generated strings. The project team will ensure that the identity of the data subject and the data about them is impossible to link together. The consent form will inform participants that their data will be shared with the public after de-identification. By implementing these measures, the project team will ensure the privacy of participants is protected, while promoting transparency and open access to research data.

- **Sustainability**

The project personnel in this multi-institutional, multidisciplinary, and collaborative study not only bring diverse backgrounds and expertise into the study, but also lead to different audiences by expanding the readers' communities. This strategic partnership among five institutions (i.e., TAMU, OU, Purdue, TXST, UNR, PVAMU) and various disciplines (i.e., information sciences, agricultural education, and education) will facilitate the widespread dissemination and adoption of project outcomes for interdisciplinary impact. Especially, researchers' affiliations with three ARL member institutions and RDAP working groups drive this study into high visibility within research data communities.

The project team will host a series of webinars introducing the measure for its use. These webinars will be promoted through library association platforms such as ACRL Online Discussion Forums and ACRL Presents. Additionally, virtual workshops will be offered to academic libraries to introduce and promote the use of the standardized measure. The standardized data core competency measure will also be made accessible to the public through the Texas Data Repository, and an open-access RDM module, based on discipline and evidence, will be made available as well.

To ensure long-term accessibility and preservation, all data and products will be stored in standard file formats (e.g., TXT, CSV, PDF). Comprehensive documentation will also support the long-term use and application of the standardized data core competency measure developed in this project. The project team is committed to providing access to these resources for a minimum of five years after the grant period, with the expectation of indefinite access in the future.

## **Data Management Plan: From assessment to implementation: Creating a standardized data competency measure and discipline-based RDM module**

### **Project Overview and Expected Data Types**

This three-year project will use a mixed-methods research approach to design a standardized measure/survey for data core competency, design a discipline-based and evidence-based research data management (RDM) module, and conduct an experimental/intervention study via pre- and post-design in graduate students' research methods courses. The goal is to equip social science graduate students (specifically, in library information science and education) with the necessary skills and knowledge to manage research data effectively and efficiently.

The project will employ a systematic approach to develop a standardized data core competency measure/survey. To achieve this, the project will conduct a systematic literature review to identify the relevant literature concerning data core competency and corresponding assessment for data core competency. Simultaneously, the project will conduct in-depth interviews with social science faculties and focus group interviews with graduate students to identify the components of data core competency. To validate the designed standardized measure/survey, the project team will collect survey data and conduct psychometric validation. This will involve item-level descriptive analysis, parallel analysis, and exploratory factor analysis. To assess the effectiveness of the designed standardized measure, the project team will design a RDM module and embed it in selected graduate-level research methods courses. An experimental study will be conducted via pre- and post- assessment, using ANOVA and regression analysis.

In the research process, the project team will produce various types of data:

- 1) research data such as surveys, one-to-one faculty interview, focus-group graduate students' interview, and experimental study data. Comma Separated Value (.csv) files will be generated to store tabular data collected from Qualtrics surveys or experimental study data. Plain Text (.txt) files and audio data (.mp3) will be generated to store interview data.
- 2) education data such as RDM module, curriculum, number of students enrolled/participating in the project, etc. PDF/A (.pdf) files and MPEG-4 (.mp4) will be generated to store course curriculum, and course materials.
- 3) research products data such as conference proceedings and publications. This project will generate conference and journal publications for dissemination purposes. All the research products will be preserved in PDF files.
- 4) digital metadata, including readme files and codebooks, will also be created for each type of primary data by the project team in (.txt) file.

### **Sensitive Information**

To protect the privacy of participants in the survey and interview process, the project will implement a de-identification process for personal information obtained in the data collection. Firstly, the project will restrict access to the raw data and only allow PI or Co-PI to access sensitive information. Next, the project will utilize pseudonymization to remove identifying information from raw data. This will involve replacing identifiable information, such as names, age, and gender, with randomly generated strings. The project team will ensure that the identity of the data subject and the data about them is impossible to link together. The consent form will inform participants that their data will be shared with the public after de-identification. By

implementing these measures, the project team will ensure the privacy of participants is protected, while promoting transparency and open access to research data.

### **Requirements and Dependencies**

To promote long-term preservation and open access to the data generated in the project, the project will use open, non-proprietary formats to store all the documents. The use of open source software such as R and RStudio will be prioritized throughout the project, and R.script will be generated and stored for all data processing and analysis. Additionally, the project will create digital metadata, including readme files and codebooks, for each type of primary data.

### **Documentation**

To ensure the research process is transparent and reproducible, the project team will capture consent agreements, data documentation, codebooks, metadata, and analytical and procedural information through the research process. All documentation will be stored in open formats, such as .csv or .txt file, and in digital format.

The published documentation will include the information necessary to read and interpret the data, such as file structures and instructions for R.script. Raw data will be preserved in a way that includes self-identifying information, which helps preserve provenance. Additionally, the project will use file naming conventions that provide a second level of visibility into the data's provenance. To ensure that the data is fully described and understood, the project team will manually generate descriptions and other metadata as appropriate.

### **Post-Project Data Management**

All quantitative project data will be shared via the Texas Data Repository (TDR). TDR provides long-term preservation of digital objects using an off-site backup and assigns a Digital Object Identifier (DOI) for citations and discoverability. By default, data is shared with a CC0 public domain dedication. Data will be accompanied by documentation, metadata, and code to facilitate reuse and provide the potential for interoperability with similar data sets. The repository provides bit-level preservation and ensures ongoing access to research data, including associated metadata and documentation for a minimum period of ten years after it is deposited. All data will be retained for a minimum of three years after the conclusion of the project or public release (publication), whichever is later. Data related to a student's research work will be retained at least three years after the student has graduated. Access to the database and associated software tools generated under the project will be available for educational and research purposes.

### **Review and Monitoring**

The PI and Co-PIs will oversee the data management plan, with updates included in progress reports and a final report that details all managed products. Intellectual property rights will be retained by the universities and investigators. Curators at the University Libraries will ensure compliance with data sharing requirements to make data FAIR (Findable, accessible, interoperable, reusable).

All the data generated from the project will be stored and shared by the TAMU team. Dr. Zhihong Xu (PI) will serve as data manager, with assistance of the graduate student in the project, and engage in quarterly quality assurance checks with team members to ensure data integrity and document the chain of custody among different datasets.