Consultant: Data Quality Improvement Plan
Request for Qualifications (RFQ)

1. Timeline for Selection Process

Activity	Date
RFQ Re-Release	January 11, 2022
Application Due	February 11, 2022
Application Review	Applications will be reviewed on a rolling basis throughout the application process.
Notifications of Acceptance/Rejection	February 25, 2022

2. Organization and Project Overview

The Child Care Alliance of Los Angeles

Background

(CCALA) is a partnership of ten Resource and Referral and Alternative Payment agencies in Los Angeles County. Together, our agencies deliver various resources and services to thousands of families and child care providers. Through our member agencies, CCALA has the ability to reach providers, families and children at a grass-roots level in multiple languages and a strong understanding of their areas' unique communities. CCALA is the glue that brings together our ten member agencies around: best practices and standardized methods in the child care voucher programs; advocacy on behalf of working parents and their children as well as in making improvements to government systems and programs; and coordination of county-wide services delivered by several or all of the agencies to families and providers in the areas of improving child care quality and well-being of the child. CCALA also manages and maintains the California Early Care and Education Workforce Registry, leads the Child Care Bridge Program in Los Angeles County, and serves as partner in the Quality Start Los Angeles Program. Visit www.ccala.net for more information.

California ECE Workforce Registry

Background

In 2012, the California Early Care and Education (ECE) Workforce Registry (herein after referred to as "Registry") began as a pilot in San Francisco and Los Angeles Counties. The Registry is an information system that collects, verifies, and tracks demographic, education, training, and employment data about the ECE workforce. In FY 13-14, First 5 LA established a strategic partnership with the Child Care Alliance of Los Angeles (CCALA) to administer the Registry in Los Angeles County.

Since then, the Registry has become a state, regional and local collaboration designed to track and promote the education, training and experience of the ECE workforce for the purpose of improving professionalism and workforce quality to positively impact children. As of December 2021, there were 120,000 Registry users across California, with 88,131 of those active users.

The Registry provides participants a single location to securely store and access:

- Verified qualifications (e.g., transcripts, permits, and other pertinent documents) that links to their employer (E-Portfolio);
- Professional development opportunities – search and enroll in trainings (Training Calendar);
- Employment opportunities – search for job opportunities in the early learning and care field and contact potential employers (Job Board and Resume Builder);
- Streamlined data used to support participation in programs such as Quality Counts California and Early Learning Workforce Pathway Grants

Since 2017, there have been several initiatives that have led to an increase in Registry profile use, including: the mandate for all California Department of Education-Early Learning and Care Division (CDE-ELCD) funded training vendors to utilize the Registry, and the integration of local Quality Rating and Improvement Systems (QRISs) with the Registry for processing staff qualifications for QRIS Tier Ratings. Currently, it is the primary source of verified data about California's ECE workforce, its professional preparation and professional development, and is a critical component of the data infrastructure for California. The Registry has over one hundred and forty professional development organizations and/or projects that utilize the Training Calendar Module to market and track training and verify attendance of the ECE workforce. The Registry is integrated with and supports staff qualification and professional development of workforce development investments for the Early Learning Workforce Pathways Grant and Quality Counts California in over forty counties, including Los Angeles, San Francisco, Santa Clara, Alameda, Nevada, with many other counties in process of integrating local efforts.

Registry Technical Manual/Codebook and Data Quality Assessment Project

Background

The Workforce Registry Codebook & Data Quality Assessment project developed a Technical Manual and Codebook. The purpose was to assist data users with requesting data from the CA Early Care and Education Workforce Registry (henceforth known as "the Registry") to conduct analyses of these data for research, evaluation and policy purposes. To facilitate analyses, Registry data must be of sufficient quality; complete, valid, and internally consistent. Part of the project included conducting a preliminary assessment of the quality of Registry data. A summary of the recommendations is included below, with the full memo included at the end of this document. The * asterisks indicate which recommendations will be pursued in the scope of work for the Data Quality Improvement Plan.

Include information from the recommendations

1. All variable names in the export files should be checked against those listed in the Codebook*
2. Provide more information to describe export data
   - When an export file is delivered to a data user, it should contain a "transmittal form" identifying criteria used to include or exclude cases
   - It would be important to know whether the file has all current/active cases or if there were other characteristics that distinguishes the cases in the export file.
3. In the longer term, Registry data should restructure multiple response variables to reduce file preparation in Excel and SPSS*
4. Reduce the amount of blank fields to increase information value*
5. Procedures should be in place to monitor levels of missing cases and make corrections as needed*

6. Create procedures to reduce missing cases*
7. To reduce the possibility of out-of-range values, the use of free-form fields for users to enter information should be curtailed so that the user interface offers fixed responses and formats for the user to enter*
    - Ensure automatic checking as data are entered to fix out-of-range values as they occur.
8. Response choices that allow users not to answer the question, such as "Do not know" or "Do not wish to voluntarily report" should be organized for all variables*
    - Group at the end of the list of valid response choices for all variables
    - Data users can identify and remove for most analyses
9. Reduce the complexity and time required to prepare data for analyses
    - To compare two or more variables from different analytic files, the files must be merged and restructured
    - Suggestion: Registry staff to perform some of the data merging prior to sending the export file to the data user
    - If possible, create export files already structured for variable matching
10. Conduct continuous monitoring of data integrity*
    - Compare verified with self-report variables, expected with actual matches.
11. Expand the comparisons of data for internal consistency
    - Identify variables and clarify assumptions regarding how they should match
    - Match variables from different units of analysis: individual, organization, event

For this RFQ, we are looking for a Consultant who has the qualifications to lead analysis and coordination of CCALA’s Registry Data Quality Improvement Plan Scope of Work, with support from CCALA’s Program Manager and Registry Director, which includes the following activities:

The consultant will:

1) Obtain orientation to the CA ECE Workforce Registry
2) Use previous analysis of the completeness and validity of existing Registry data, and recommendations for improvement (Data Quality Assessment)
3) Develop an approach to the Data Quality Improvement Plan
4) Develop a Data Quality Improvement Plan
5) Implement the Data Quality Improvement Plan
6) Develop a mechanism to track progress in implementing the plan
7) Memo/Summary that explains the status and next steps

3. **RFQ Overview**: The Child Care Alliance of Los Angeles (CCALA) is seeking qualified consultant applicants to coordinate the planning, development and implementation of CCALA's Scope of Work for Data Quality Improvement Plan. The services coordinated and developed through this grant will support an aligned, systematic, state-wide effort to increase the data quality in the California Early Care & Education Workforce Registry to improve access to critical data for the State, counties, regions, employer, individuals, policy makers and researchers.

   a. **Contracted Consultant**

      i. **Project Term:** February 2022 – December 31, 2022

      ii. **Not to Exceed:** This solicitation is being issued with a not-to-exceed amount of $75,000 for the six-month duration of the contract. The Consultant will be required to construct a deliverables-based budget for the six-month period to cover all planned tasks with the contract year.

      iii. **Consultant Rate**: The total composite rate for a Consultant may not exceed $150 an hour. This means that the total cost of billable hours associated with a contract divided by the total number of hours billed must be equal to or less than $150. A blended rate is allowable. For example, for a contract totaling $15,000, a consultant may bill 50 hours for Consultant A at $200/hour, and 50 hours for Consultant B at $100/hour, with a total composite rate of $150/hour (100 total hours divided by $15,000 in billable hours = $150/hour).

      iv. **Location:** Virtual/Remote with some in-person meetings if needed and in accordance with the Center for Disease Control (CDC) guidelines and agency policies. Any in-person work would be within Los Angeles/Placer County. No travel expenses will be reimbursable.

Eligibility: Eligible applicants must demonstrate the qualifications, experience, competency, and ability to successfully lead the coordination of Data Quality Improvement Plan grant efforts, as indicated. Applicants must submit their application under their legal name or legal business name (if applicable).
Required

Qualifications

- Expertise in analyzing qualitative, quantitative, and administrative data
- Experience in conducting psychometric analysis
- Experience in designing and conducting interviews
- Experience assessing data quality include validity and completeness
- Experience in designing clear and informative user manuals and codebooks for datasets or data systems
- Skilled in MS Excel, Access and other relational database software
- Skilled working with databases and data structures
- Excellent verbal and written communication
- Demonstrated capacity to complete tasks and deliverables, and manage complex data projects
- Knowledgeable about data privacy and confidentiality issues especially in the context of administrative data
- Experience with technology to recommend how data could be captured differently in the Registry (from a software perspective)

Preferred Qualifications

- Familiar with early care and education data and policy issues
- Excellent planning and research skills
- Excellent critical thinking
- Excellent organization skills

- Flexibility, adaptability and integrity

b. **Desired Competencies:**
   i. **Diversity:** Shows respect and sensitivity for cultural differences, actively works towards equity, access and inclusion in all aspects of work
   ii. **Ethics:** Treats people with respect; Keeps commitments; Inspires the trust of others; Works with integrity and ethically
   iii. **Project Coordination** - Develops project plans; Coordinates projects; Communicates changes and progress; Completes projects on time and budget; Coordinates project team activities
   iv. **Adaptability:** Adapts to changes in the work or work environment; Manages competing demands; Changes approach or method to best fit the situation; Able to deal with frequent change, delays, or unexpected events.
   v. **Planning/Organizing** - Prioritizes and plans work activities; Uses time efficiently; Plans for additional resources; Sets goals and objectives

4. **Terms of Participation:**
   a. **Acceptance of Qualifications does not guarantee a contract with CCALA.** The selected consultant and CCALA staff will negotiate a scope of work and final budget during the contracting process. Any performance or services begun prior to receiving all written approvals by CCALA shall be considered voluntary and are not subject to reimbursement.
   b. Initial contact and on-going correspondence will be done via e-mail. Applicants are responsible for providing a valid e-mail address during the application process and notifying CCALA of any changes during the term of the contract. If a valid e-mail address is not on record, CCALA may determine that the applicant is no longer able to complete the work and cancel the contract.
   c. The term of participation for project is from the time of acceptance through December 31, 2022.
   d. Consultants may withdraw their application at any time by e-mailing a signed letter to CCALA. Accepted consultants are not bound to accept the work solicited by CCALA.
   e. CCALA reserves the right to amend the qualification requirements as needed to best meet the needs of all parties. At CCALA's discretion, consultants may be removed from consideration at any time.

5. **Selection Process and Review Criteria**

CCALA will review the applicants based on the following multi-phase review process:

<u>Phase 1:</u>

CCALA will review all applications for completeness and minimum requirements. Basic requirements include: timely receipt of application, submission of all required attachments, etc. Applications with omissions of any required documentation are subject to disqualification.

<u>Phase 2:</u>

Applicants that pass Phase 1 review will proceed to Phase 2 review.

***The Consultant must provide a response to this solicitation in order to be considered for the contract. Interested consultants must submit the following*:**

6. Required Documents

Documents required to apply to this RFQ:

A. **On-line Application**: RFQ On-Line Application
Note: Documents B-G will be submitted through the RFQ On-line Application link above.

B. **Proposal Narrative**: The narrative should be single-spaced, 11-point font with 1" margins and should not exceed 3-pages. Information beyond the page limit will not be considered.
C. **Expertise**: Discuss your knowledge, skills, and experience related to working with data systems, analyzing and assessing data quality, and developing data quality improvement plans. Include 2 brief examples.
D. **Approach**: Provide a general proposed approach for completing the tasks and accomplishing the objectives of this project.
E. **Qualifications**: Submission of resume demonstrating required qualifications. Include resumes or curricula vitae for all proposed personnel who will exercise a major role in carrying out project tasks. Resumes must not exceed five (5) pages.
F. **Budget/Pricing**: Submission of budget/pricing.
G. **References**: List three (3) current or former references where similar work was performed or accomplished. Please note that CCALA reserves the right to contact references in order to determine fit for the proposed work.

Failure to submit all required attachments will be considered an incomplete application and may result in disqualification from the process. In order to respond to this RFQ, please complete and submit your on-line application and all required documents to CCALA no later than **February 11, 2022**. Applications received after this deadline will not be considered.

7. Terms of the RFQ

CCALA reserves the right to reject all applications and re-solicit for this RFQ. Failure to comply with application requirements shall be grounds for disqualification.

CCALA shall not be liable for any costs incurred in connection with an applicant's preparation of an application in response to this RFQ. Any cover letters, narrative, resumes, and curriculum vita, including attached materials, submitted in response to this RFQ shall become CCALA's property and subject to public disclosure.

The applicant agrees that, by submitting an application, the applicant authorizes CCALA to verify any or all information and/or references submitted in the application.

All materials developed as a result of this RFQ will be open-source and available to the public.

8. Document Submission

Only complete submissions will be reviewed. All documents should be submitted through the application link above. If you have any questions about or issues with document submission, please contact: Elise Crane, Registry Director at Elise.Crane@ccala.net or Fiona Stewart, Program Director at Fiona.Stewart@ccala.net.

Workforce Registry Codebook & Data Quality Assessment Project
Data Quality Assessment Memorandum

Gary Resnick, Ph.D.

Draft #4: March 25, 2021

I. Introduction

The Workforce Registry Codebook & Data Quality Assessment project developed a Technical Manual and Codebook, which includes a Data Request Form. The purpose was to assist data users with requesting data from the CA Early Care and Education Workforce Registry (henceforth known as "the Registry") to conduct analyses of these data for research, evaluation and policy purposes. To facilitate analyses, Registry data must be of sufficient quality; complete, valid, and internally consistent. Thus, it was necessary as part of this project to conduct a preliminary assessment of the quality of Registry data.

This memorandum describes the results of that assessment using existing quality standards for completeness, validity, and integrity and provides recommendations for improving Registry data quality. The standards developed by the California Department of Education (CDE) for all large-scale administrative data systems[1] was followed for this Data Quality Assessment (DQA). High-quality means that the Registry data are complete, consistent and reliable. By adhering to a set of established standards, data users can conduct analyses to help advance ECE practice, quality improvement, research and policy.

Figure 1 gives a graphical depiction of the DQA process. First, a request for Registry data was made, based on identifying approximately 10-15 variables from the Codebook at each of the three units of analysis (described more fully in Appendix A). As noted in the "Limitations" section of this memorandum, since the request form is still being revised, it was not used to make the request, as it would be eventually employed by data users.

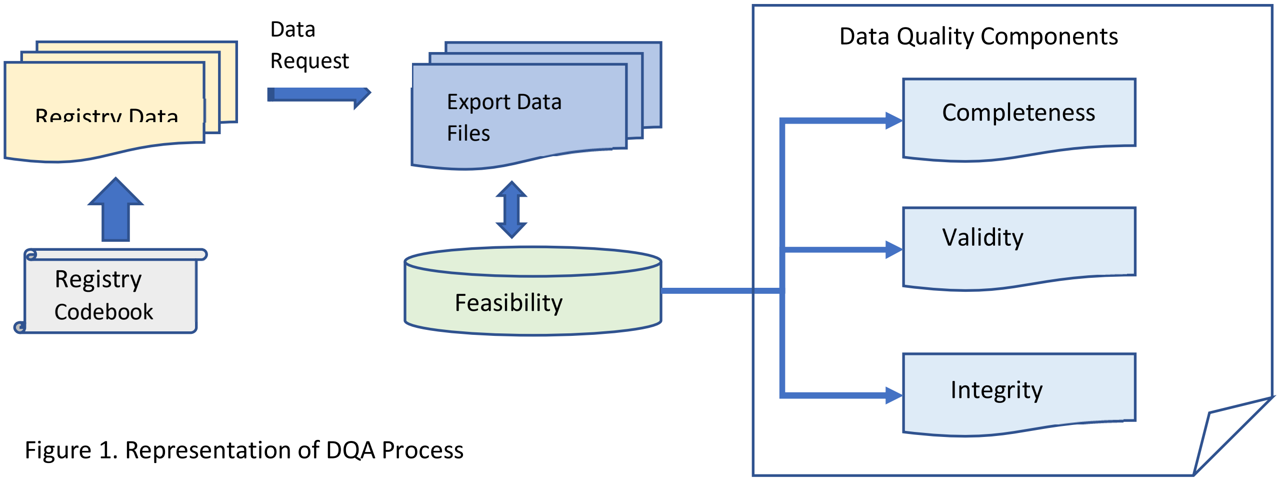

Figure 1. Representation of DQA Process

Second, the Registry office sent a number of data files in Excel formats and, third, a set of data preparation procedures were conducted prior to analyses of data quality. These first three steps informed the first aspect of data quality, feasibility and served as a form of pilot test of how this process will operate for actual data users, except without use of a Data Request Form, which is still under development. Finally,

[1] Data Standards and Inventory. The Educational Data Governance Program (EDGO). CA Dept. of Education. January, 2019. Accessed at: https://www.cde.ca.gov/ds/ed/datastandardinventory.asp

the three dimensions of data quality – completeness, validity and integrity - will be assessed using primarily descriptive frequency analyses.

The complete set of processes for conducting the DQA are described in Appendix A and summarized in the Sidebar "Steps in the DQA Process." The key findings, will be summarized for each of the key dimensions for assessing data quality, including feasibility, completeness, validity and integrity. Recommendations for improving Registry data quality will be given in order to initiate further discussion and planning. All analytic tables cited in this memorandum can be found in Appendix B "Supporting Tables." Supporting output from the statistical analyses is available upon request.

II. Results: Feasibility

Background. Feasibility is defined in the Cambridge Dictionary as "the possibility that can be made, done, or achieved, or is reasonable."[2] In terms of this study, the DQA assessed whether a set of procedures can work to provide data for descriptive and statistical analyses. For example, the study initially assessed the extent to which it is possible to select a set of variables from the Codebook, request these variables, receive one or more export data files that were checked against the Codebook, and employ these files for statistical analyses. It would also be important to anticipate potential problems that might be faced by data users in obtaining or working with a data export file once it is received. Ultimately, the feasibility of procedures expected to be employed by data users will determine both the quality of the data and whether, and how, they can be analyzed.

Steps in the DQA Process

Step 1: Request Registry data

Step 2: Examine Excel export files

Step 3: Prepare Excel files for import into the Statistical Package (SPSS)

Step 4: Conduct descriptive analyses using SPSS Statistical Package

Step 5: Conduct Internal Consistency Assessment

Results. Three key challenges arose when feasibility was assessed. The following results will be discussed according to these three challenges, including: a) inconsistencies in the data export files received, b) missing requested variables from the export files, and c) variable formats in the export files did not match those specified in the Codebook.

Inconsistencies in the Data Export Files Received. A request was made to the Registry office to obtain approximately 10-15 variables[3] from each of the three levels (individual, organizational, and event). The request asked that the data be sent via three Excel export files, corresponding to each of the three units of analysis. Variables selected for each module included those from a variety of formats (numerical, text, or dates) and those that were both self-reported and verified. Finally, to the extent possible, variables were selected based on whether their values would be expected to match, such as highest educational attainment and degree level.

The original request for three separate export files resulted in the delivery of eight export files, several at each of the three levels of analysis (individual, organization, and event). Table 1 summarizes key descriptive information about the eight export files. Overall, 40 variables were produced in the eight export files across the three level of analysis. For each export file, the number of records contained is also listed in

[2] Accessed at: https://dictionary.cambridge.org/dictionary/english/feasibility

[3] In this memorandum, "variable" will be used instead of "field" to identify each data item in the database. For each variable, the Codebook specifies the range of response values, and labels, assigned. Please refer to the Codebook for more information.

Table 1. Also, as per the initial request, all of the export files included **UserID** or, in some cases **ProgramID**, as the match key variables, in order to check for duplicates and to facilitate the merging of files later.[4]

Missing Requested Variables. When comparing the list of variables requested with those received, there were a number of variables listed in the Codebook that were requested but not received, as follows:

- **ECERole:** Role in the Field
- **AgesAppliedTo:** Child Age Groups Training
- **PrimaryCKAID:** Primary Core Knowledge Areas

Variable Formats in Export Files did not Match Codebook: Most variables in the export files were not in the formats specified by the Codebook, whether fixed numerical codes, continuous numbers, or date formats. The Excel files were primarily formatted as text strings, even if the Codebook specified a set of fixed response choices with numerical codes. As a result, extensive formatting in Excel and, in some cases, the SPSS statistical software, was required before each dataset was ready for analysis. With regards to variable names, most variable names in the export files were identical to those requested via the Codebook, except for ten of the 40 variables where variable names in the export files were either spelled differently from those listed in the Codebook or, in one case, a variable in the export file was not included in the Codebook. Table 2 lists the variables where the names in the export files did not match those in the Codebook.

Assessment Conclusions. In general, it possible for a Registry dataset to be exported based on a data request, meeting the definition of feasibility, but there are a number of key problems data users may encounter in obtaining the data files that match to the Codebook. The inconsistencies between the number of export files requested and what was ultimately delivered will likely be confusing to data users. As well, there was insufficient information provided to explain what each of the export files comprised, nor why eight separate files were required.

As well, some variables were missing from the export files, and a significant number of variable names in the export file did not match those in the Codebook, or the variable formats did not match with those in the Codebook. These last two issues are important because data users will be expected to consult the Codebook in order to make a data request.

Once received, the inconsistencies in variable names and formats required extensive dataset preparation before the Excel files could be imported into a statistical package. Some of these steps were relatively simple, such as formatting the variables appropriate to their use as specified in the Codebook, rather than receiving variables formatted as text strings. But other steps required more complex preparation, such as dealing with multiple response variables (explained in Appendix A). Further, once the statistical files were created, additional restructuring was usually required, in order to merge each of the small, disparate files at the same data level (e.g., user) before a single, analytic file can be available for analyses. In the future, there should be procedures in place for ensuring that the Codebook more closely fits with the export data files, and vice-versa, with approaches for maintaining consistency as the Registry database continues to be updated.

Overall, feasibility seems to be relatively low at this point and leaves significant room for improvement in the future. However, this view should also be placed within the larger perspective that the Registry is an administrative database and was not designed as a public use dataset. The Codebook developed in this project was meant to reflect the existing state of the Registry database by summarizing key attributes of all Registry fields. No such complete and up-to-date document existed prior to the development of the Codebook. The existing state of the Registry database revealed that the database fields

[4] **ProgramID** served as the match key for organization level data.

were not sufficiently specified to meet the conventional standards for a Codebook. The task for the future should focus on improving consistency and standardization between the Registry database and the Codebook, in order for the Codebook to act as an up-to-date reference source for data users.

Recommendation 1: All variable names in the export files should be checked against those listed in the Codebook. For example the variable names used in the export file should be checked against those listed in the Codebook so that there are identical names and formatting consistent with that in the Codebook. Systems and processes need to be in place to ensure the consistency is maintained as the Registry database is updated.

Recommendation 2: The Registry office should provide more information to describe export data when fulfilling a data request. For example, when an export file is delivered to a data user, it should contain a “transmittal form” identifying criteria used to include or exclude cases. It would be important to know whether the export files contain all current/active cases or if there were other characteristics that distinguishes the cases in the export files sent to data users.

Recommendation 3: The Registry database should be restructured to handle multiple response variables, in order to reduce file preparation in Excel and SPSS prior to statistical analyses.

III. Results: Completeness

Background. Completeness refers to the amount of missing cases for variables present in the Registry dataset. Large numbers of blank or missing cases can negatively affect the confidence one has in the accuracy of the data. Missing cases may not always be evenly distributed across a data file. The higher numbers of missing data for particular sub-groups of participants, organizations or events can also produce systematic biases and reduce the generalizability of results. Further, large numbers of missing cases can also make integration of the Registry with other databases more difficult.

Another aspect of completeness is whether the cases in the export file represent all possible cases in the Registry file. This concept, termed comprehensiveness, is a key limitation of the DQA because no information was included with the export files as to whether they included all possible cases or not (see "Limitations"). In this section, below, one example of potential problems with comprehensiveness is described but, overall, the DQA did not have data to assess comprehensiveness.

By conducting descriptive frequency analyses, the DQA identified different types of missing values, including those that are left blank when a value should have been entered, as well as those that are blank due to a "Not Applicable" response. Missing response values can also occur when the user does not know or have the information that is asked for, or refuses to answer. In the "Validity" section that follows this section, the issue of the order in which non-valid responses were grouped among the list of valid response codes will be discussed, but here the issue is that there are no specific response choices for many variables, allowing for blanks rather than the type of non-valid response choices. In some cases, it would be important to know how many users gave a "Don't Know" response since this suggests the user cannot enter the data requested while, in other cases, "Not Applicable" responses can be checked for correctness of the skip pattern. Thus, there are important reasons to code and record these different types of non-valid values rather than leaving them all as blanks, which appears to be the current practice in the Registry database for most variables.

Descriptive frequency distributions were examined for variables from all eight export files to identify missing cases. Completeness was assessed by looking at the percentage of missing cases from valid cases, for each variable.

Results. The large majority of variables in the eight export files contained few or no missing cases, indicating a high level of completeness (Table 3). The only caveat is that, for most variables, we do not know if the cases in the export file represent all possible cases in the Registry file because no information was included with the export files (see "Limitations"). Only two of the 40 variables indicated percentages of missing cases that exceeded the conventional standard of 10%.

Some variables had larger percentages of missing cases due to skip patterns, that is, where users legitimately skipped fields due to responses given in an earlier field, or where the field is optional. For example, 82.6% of cases for **Approved_Units**, in the *stipend* export file, were missing, but valid cases were only applicable for those applications that were approved, which was a subset of all cases in the export file. When calculating frequency distributions, these cases would typically be identified as missing. For **CKA** (Instructor Core Knowledge Areas), from the *instructor* export file, 27.1% of cases were missing. But this is explainable by the fact that this field is optional for instructors to enter information. Thus, the high number of missing values is indicative of instructors who chose not to enter this information.[5]

In the case of **OccupationCompensationType,** from the *user program* export file, 22.9% of cases were missing, yet two related variables, **OccupationHoursPerWeek** and **OccupationCompensationRate** from the same *user* program export file, had far fewer missing cases. In theory, the numbers should be similar for all three variables since they refer to different aspects of employee compensation, yet they are not. Most interesting in this case is that the actual compensation rate in dollars, which may be considered by some as more sensitive information, did not have the greatest number of missing cases, so we can rule out refusal due to users' sensitivity of the information they are asked to enter as a possible reason for the number of missing cases. It seems more likely that there is some aspect of how type of compensation is entered in the user interface where perhaps it may be easy to accidentally skip, and this should be investigated further.

As noted above, another aspect of completeness is whether the cases in the export file represent all possible cases in the Registry file. This concept, termed comprehensiveness, was a key limitation of the DQA because no information was included with the export files as to whether they included all possible cases or not (see "Limitations"). The degree to which the original export files from the Registry contained all valid records cannot be determined. This issue is discussed more fully in the "Limitations" section.

Conclusions. A strong aspect of the Registry data quality, most variables exported for the DQA had few or no missing cases, suggesting a high level of completeness. For some variables, the high percentages of missing were due to non-applicable cases from skip patterns and thus would not be problematic. For one variable, **CKA**, the large number of missing values was due to the field being optional for instructors. However, since this information is optional, the value analytically for data users is minimal since there are so many missing values. Also, we do not know from blank cases whether the instructor chose not to report or whether the instructor missed the field. In the future, consideration should be given towards requiring instructors to enter these data and this also points to using meaningful codes rather than blanks to understand why the data are blank for a given case.

Finally, it should be noted that while almost complete data were available for **Degreeawarddate** in the *user degrees* export file in Excel. However, when this variable was imported into SPSS, 81% of the cases became considered missing. Upon further investigation, it was discovered that most of the data were entered only as text strings and were not all using the same date format. When a date format is declared in

[5] As an aside, it is an open question as to why instructor information was considered optional, which perhaps should be discussed. The role and career pathways of ECE instructors should be important topics to investigate analytically, but this cannot happen if data entry is optional. We do not know what factors might differentiate those who enter such information from those that do not, thereby producing a potentially biased analytic dataset on instructors.

SPSS for any given variable, such as the conventional *mm/dd/yyyy*, then all cases for that variable will be transformed from string to that date format, as long as the data are in the same string format so that the computer can make the conversion. This happened for 1855 cases. For most of the remaining 7900 cases,[6] the statistical program could not convert the data into the correct date format and thus were assigned as "system missing." For the purposes of this analysis, Table 3 indicates that, for this variable, there were few missing cases. But, for dates to be used analytically (e.g., to compute time between dates), all cases must be using the correct format.

This issue highlights the more general problem in which most data in the Excel files delivered for this DQA were formatted as text strings rather applying the correct format for a given variable to be used analytically, such as the date format for variables that are clearly dates. Not only must the same format be used for a given variable, all cases with data must consistently apply the same format, especially for dates where there are a variety of ways for representing them. In the future, all date formats exported to an Excel file should use the same type of date format consistently, and this applies to formats of other variables referenced in the Codebook. Revising existing data in the database requires additional error checking to ensure that all cases will be consistently in the same format. One of the sources of this problem, as discussed earlier, is the user interface in which the data must initially be entered in the format that will be used for analysis, rather than allowing for an open-ended text string where users can enter whatever format they wish (despite instructions to put it in a given format) and errors can sometimes occur if blank spaces were accidentally added making text strings unreadable as dates. While the handling of string-formatted data may be more acute for SPSS than for other statistical programs, in general no statistical program can handle variables where individual cases have inconsistent formats.

To summarize, the key challenges for data completeness, there is a lack of standardization of the user interface along with often vague variable definitions and ranges of values. For example, limits on response ranges, establishment of mainly closed-ended responses while reducing free-form entry, and alerts to users when fields are left blank. Also, specific values were not assigned for different types of blank responses especially when there are legitimate skip patterns, such as in the case of "Not Applicable" and "Don't Know" responses. As a result, blanks in variables are not coded by different reasons, which could provide additional information for data users. These issues produced to the following recommendations for improving the completeness of Registry data.

Recommendation 4: Reduce the amount of blank fields to increase information value. For variables with skip patterns faced by users for whom the variable does not apply, rather than leave these blank, unique values should be assigned as "Not Applicable" so that these can be distinguished from fields that are left blank, either by mistake or due to refusing to provide the information. As well, all blanks should be examined for other forms of missing values, such as "Don't Know" or "Don't have this information" as appropriate and assigned unique values for all variables, so that blanks would remain only for those fields that are accidentally left blank by the user.

Recommendation 5: Procedures should be in place to monitor levels of missing cases and make corrections as needed so that whenever a data request comes in, there is confidence that the export files contain complete data.

[6] Additionally, there were another 7 cases that were problematic. Four did not have sufficient information ("1969," "1971, 1989," and "September, 1988") while the remaining three were in the wrong date format (e.g., "2014/05/14," "2014/05/14," "6/30/2006," "8/31/2013") and were thus declared missing.

Recommendation 6: Registry staff should create procedures for the ongoing monitoring of data comprehensiveness. Variables included in export files should represent all current and active cases, unless otherwise specified by the data user, and information given about whether the export files reflect only complete and active profiles or not, with clear definitions for complete data.

IV. Results: Validity

Background. Data validity means that each variable in the Registry has a consistent set of valid response values, as defined in the Codebook, so that the data export files entered do not contain different response values that were not specified, known as out-of-range values. This definition focuses attention on the *allowed or valid response choices* that users are given when entering their data. The DQA examined the prevalence of out-of-range values for a sample set of variables from several exported Registry data files. All out-of-range values refer to situations where the actual responses for each variable that appeared in the export files did not match the ranges of the values expected, based on the specifications in the Codebook. There can be different reasons for why unexpected response choices occurred (discussed below). Patterns of out-of-range values for specific variables may be due to the data entry interface, in which instructions to users about what to enter may be unclear, but may also be due to insufficient error checking procedures, either at the user interface or at the back end of the database (once data were entered). When these unexpected values or ranges of response values appear in the export files that will be sent to data users, it is likely that the data user will not know whether or not these are valid values and will require more data preparation before they can conduct statistical analyses.

Results. Based on an examination of the frequency distributions, most variables in each export file had appropriate ranges of response values matching the Codebook specifications. However, 18 of the 40 variables in the export files revealed a number of problems with a variety of types of out-of-range values, where response values do not match with the Codebook or differed in other ways from the Codebook in how the response values were labeled or formatted (see Table 4). The types of problems covered under "out-of-range" that will be summarized in this section include the following:

a. Export files that contain values not in the Codebook;
b. Differences in how valid values were defined or ordered in the export file;
c. Formatting of export files that did not match formats specified in Codebook; or
d. Inconsistent text string formatting.

In addition to describing the results by giving example of each type, this section also discusses a related problem since it refers to how response choices can be summarized and described, that is, the issue of multiple response variables and how they are structured in the Registry dataset.

Export files that contain values not in the Codebook In the *training events* export file, **eventCapacity** is a free-form entry for the number of individuals that may enroll in the event yet some values appear either too large (e.g., "1000," "1999," "9999," "10000," and "99999") or too small (e.g., "-3") when compared to the specification in the Codebook. This is an example of potential problems with the user interface where it is not clear to the user how they should enter information, particularly if the field is free-form, where any number or date is acceptable.

Differences in how valid values were defined or ordered in the export file. Variables where users have multiple ways of not answering, or in some cases, refusing to answer, the question, such as leaving the response blank, or responses. For example, the variable **userSecondaryLanguage**, non-valid response choices mixed with valid response choices included "Do not wish to voluntarily report," "Did not answer," and "None" (with no explanation of what this means). This makes it difficult for data users to identify and remove these responses in order to obtain frequency distributions for valid response. Further, for **StipendPathGoals**, only 14 of the 17 possible response categories were selected by users. The responses "completed_esl," "high_school_diploma," "completed_ged," listed in the Codebook for this variable were not selected by any users, suggesting these categories may not be needed.

Formatting of export files that did not match formats specified in Codebook. In this situation, the data in the export files were not formatted according to those specified in the Codebook. For example, in the variable **DegreeAwardDate** some dates in the export file used the *dd-mm-yyyy* format (e.g., "22-May-2009"), whereas the convention for date formats in the United States, and as indicated in the Codebook, was in the *mm-dd-yyyy* format. But, more importantly, not all cases for this variable were formatted in the same way. As a result, when this variable was imported into SPSS, 81% of the cases became considered missing or invalid. When a date format is declared in SPSS for any given variable, such the conventional *mm/dd/yyyy*, then all cases in that variable will be transformed from string to that date format, as long as the data are in that format so that the computer can make the conversion. This happened for 1855 cases. For most of the remaining cases (N=7900), the statistical program could not convert the data into the correct date format and thus were assigned as “system missing.” For the purposes of this analysis, these cases were not considered as missing (see Table 4). For dates to be used analytically (e.g., to compute time between dates) all cases must have the correct format. The inconsistency of data formats for cases in the same field suggests the need for greater standardization for the user interface, including specific date formats with range checks or warnings.

Inconsistent text string formatting. The majority of variables in the eight export files were formatted as text strings, which is usually used for open-ended responses but not for numerical codes or other specific formats, such as currency, dates, etc. The problem with text strings is that any manner of spelling of responses could be valid, even though statistical programs handle these formats as if they are different response choices, such as the variable **Semester** in which the Fall semester is spelled both as "FALL" (upper-case) and "Fall" and thus are treated by a statistical program as different responses. The Codebook specifies that the valid text should be "Fall." The use of text strings is among one of the most typical and yet the thorniest of issues when conducting statistical analyses, particularly in SPSS, because every change in spelling is considered by the statistical program as a different variable, and also because response choices require numeric codes for most statistical analyses. This issue is discussed further in Appendix A and best practices suggests that text string formats in export files should be minimized.

Multiple Response Variables Not Structured for Analyses. Another issue related generally to interpreting response choices for variables in the export file, but that is not considered an issue with out-of-range values, occurs with the use of multiple response types of variables described in detail in Appendix A. The main issue is that responses have to be handled differently than variables where the response choices are mutually exclusive (only one response can be selected). When users are able to give more than one response, the number of responses will not match the total number of cases, since there will be more responses than cases. As a result, starting at the data entry point, the database must be structured to handle multiple responses. For most of these variables in the Registry dataset, the structure is inadequate to enable these variables to be properly analyzed (discussed further in Appendix A).

For example, Table 5 lists all potential response values for **StipendPathGoals**, and displays how frequency distribution output should appear for multiple response types of variables. This table was produced after extensive data preparation of the initial Excel file and the resulting SPSS file. As the table revealed, the response categories were summed according to both the total number of all responses, N=13,976, and the total number of individuals, N=10,279. However, since each individual may have completed more than one application or selected more than one goal per application submitted, it would be misleading to base the percentage distribution on the total number of individuals (for most study questions). When reporting percentages from multiple response variables, the total number of responses, rather than the number of cases/individuals, should be reported.

Conclusions. For most variables, the export file matched the Codebook for the response value labels shown in the frequency distributions, but there were a number of notable exceptions. Out-of-range values were found for a number of variables in each of the export files (most notably in the *Stipends* export file) when compared to the set of valid response values listed in the Codebook. It seems that many of these problems may have occurred at the front-end of the database when users were entering data. When fields are completely open by using text string formats that allowed for any responses to be entered, no alert or warning could be given to users about whether what they entered is what was expected. With closed-ended fields offering a specific list of responses, error checking should alert users if they try to enter out-of-range response. Finally, the lack of grouping of all non-valid response choices within the list of valid choices does not conform to best practices for these no-information values; instead, these response categories should usually be grouped either at the beginning or end of the list of valid values to help data users more easily identify all types of non-valid responses. Much more standardization and guidance for users is needed to reduce or eliminate out-of-range values.

Recommendation 7: To reduce the possibility of out-of-range values, the use of free-form fields for users to enter information should be curtailed so that the user interface offers fixed responses and formats for the user to enter. Also, procedures should be in place to ensure automatic checking as data are entered to fix out-of-range values as they occur.

Recommendation 8: Response choices that allow users not to answer the question, such as "Do not know" or "Do not wish to voluntarily report" should be organized for all variables. For example, non-valid responses should be grouped at the end of the list of valid response choices for all variables, allowing data users to identify and exclude these values for most analyses.

V. Results: Integrity

Background. Integrity is defined as the internal consistency of the data. Internal consistency is applicable to all records, regardless of when and where information is entered or who enters the data. Data that are internally consistent will reflect clear-cut standards for data entry and verification in the Registry. These analyses assessed the degree of consistency between two sets of variable pairs, employing an exploratory approach. It should be expected that user-entered education level would be matched with verified user's degree level from the educational institution database. It should also be expected that primary and secondary languages the participant spoke should match with those languages spoken to children in the program. In both cases, two sets of variables that should be consistent will be matched. One caveat; for this exploratory approach, the assessment of integrity in the DQA was limited to Registry data at the individual level of analysis and involved comparing the consistency of two similar variables for two types of such variables (in the latter case of languages, three variables were compared).

Results. In order to assess integrity, it was necessary to merge data from several of the eight export files that were delivered. The processes for merging the data are described in Appendix A. In general, results of this merge indicated that, as additional files were added incrementally to the initial and largest export file, the user export file, the number of cases (Ns) with valid **UserID**s fluctuated. Each time files were merged, the numbers tended to slightly increase, with the largest increase occurring when the *stipends* file was added to the merged file (N=65). These findings suggest that some users may have had data only in the *stipends* file and nowhere else, which should be investigated. This section describes the results of matching each of two sets of variables from the merged individual-level dataset.

Matching Educational Levels. Table 6 shows the degree to which the four main educational categories matched between two similar variables, with an overall match percentage of 77.4% across four categories of education levels, which suggests a very high degree of data integrity.[7] Self-reported **highestEducationLevel** from the *user* export file contained 102,124 cases, and this was matched to **DegreeLevel** from the user degrees export file with 9,762 cases.

It should be noted that there is no established standard for considering integrity as being "high," but the fact that both variables needed to have valid values otherwise the case would be dropped from the analysis, reducing the sample size and hence statistical power, and that the match involved the same, specific categories for each variable, suggests that 77.4% is indicative of high integrity. Further, it is possible that any mismatches may not necessarily represent a data quality issue but possibly may indicate differences between verified and self-reported data sources.

Table 6 also revealed that the match percentages were lowest for users at the Associate's level (50.4%), both when this category included Some College or not.[8] Those with higher education levels generally revealed higher percentages of cases that matched (92.9% and 91.5% for Masters and Doctorate, respectively), and 74.6% for those with a Bachelor's degree.

Matching Languages. Three separate analyses were conducted to match each pair of **userPrimaryLanguage, userSecondaryLanguage**, from the *user* export file and, **ProgramLanguages**, from the *user program* export file. Overall, 102,016 cases matched primary languages with secondary languages, while 2,568 cases matched primary languages with program languages, and 3,975 matched secondary languages with program languages.

Table 7 displays a high match percentage between users' primary language and the secondary languages that users reported, averaging 70% across all languages. Interestingly, English was not the language with the highest match percentage (80.7%) and neither was Spanish. Also, as expected, the "catch-all" category of "Other" was among the lowest matched languages of the pair.

Table 8 reveals that the average match between primary language and program languages was considerably lower at 40.8%. Finally, when comparing secondary languages and program languages (Table 9), the average match percentage of 45.2% is slightly higher than that comparing primary language and program languages. The match between secondary languages and program languages was high for English, since this would be expected to be the language of instruction, but it was also highest for speakers of Hmong and Farsi.

Conclusions. There are important caveats when interpreting these matches. First, these analyses were exploratory, using only two variables and six matches overall. Second, there could be a number of valid and varying reasons for finding low match percentages that may not indicate a lack of consistency.

[7] For the purpose of this analysis, we combined the “Some College” and “Associates” for **highestEducationLevel**. When the files were merged, we identified 9,740 cases with information from both files. There are several reasons for doing combining these categories: a) the Associate's category holds a significant number of cases that would otherwise be omitted, b) if the education level categories are ordered from lowest to highest, Some College *follows* the Associate's category, suggesting it is at a higher level, and would appear to users in this order, and c) it is a conservative adjustment because ordinarily Some College would indicate a higher educational level relative to Associate's. Thus, matching highest educational level to degree level included the "Some College" combined with the Associate's category from the highest educational level. If Some College was removed, the overall match percentage would be slightly lower, at 76.4% but overall the results remained consistent.

[8] As a check on the match percentage for each category, these analyses were redone excluding Some College and solely with the original Associate's category, without Some College. The match percentage for only the Associate's category was lower (46.7% compared with 50.4%).

These analyses cannot identify the reasons for a low match, particularly since, especially in the case of languages, there were many languages listed by users. The greater the number of categories to match, the decreased likelihood there is a finding a match. Also, we cannot assume or even expect that language(s) spoken or a primary language should match program languages for a host of reasons and may not be indicative of low data integrity. Finally, , no benchmark or threshold currently exists for what would be expected to indicate high consistency; there is no objective definition for a high match. Given these caveats, we would still expect some degree of consistency between the pairs of variables that were compared, particularly for the educational variables.

The internal consistency of the data as an indicator of data integrity, based on matching two variables from each of two different content areas (education level and language), appears to be very good. The overall percentage for two similar educational variables measuring highest educational degree achieved, one self-report and the other verified, indicates a high match. This is notable because very specific categories were matched. The match percentages for each valid response value were lowest for users at the Associate's level whereas those with higher education levels generally revealed higher percentages of cases matched. The lower percentage of matches for the Associate's level may be due to possibility that information at this level may comprise more varied educational experiences reported by users. One possibility to investigate is that the verified Educational Institution database may have better information at the higher degree levels than it does for the Associate's level, which could comprise a variety of different educational experiences and degrees. It is also possible that the user interface may have made it difficult for users to select the correct category. These are only some of the potential hypotheses that should be pursued in the future.

The matching of three language variables was complex and can be difficult to interpret. Overall, For many languages, the language matches in the tables (Tables 6-8 in Appendix B) appear relatively high, particularly for the match between users' primary language and the secondary languages that users reported. Matches between languages spoken and those spoken to children at an ECE program were lower but there was no strong match, except English and Spanish, where one's primary language was also one of the languages spoken in the program. The lower percentages of some matches may be due to the relatively few languages spoken in most early care and education programs combined with the large number of languages that were compared for each variable and the relatively low number of cases for many of these languages. Thus, the chances of finding a match would be relatively low and, for this reason alone, we can say that the match percentages appear relatively high, even if they seem low in absolute terms. It is also important to be careful interpreting these matches because there could be valid and varying reasons for match percentages that are relatively low and may not necessarily indicate a potential lack of consistency.

Additionally, matches between languages spoken and those spoken to children at an ECE program may also reflect key characteristics of the underlying workforce population in which English is often a second language. For example, a significant demographic (although perhaps not based on absolute numbers of cases) comprise Hmong and Farsi speakers, reflecting larger percentages of these speakers working in ECE programs. These are speculations and, as mentioned above, there is no clear-cut reason why the ECE workforce would match some languages with the program.

Only four pairs of variables (eight variables in total across two content areas) were matched, due to time constraints, and these were done only at the user level. Future analyses of Registry data should extend and routinize consistency checks with variables that are expected to match, and go beyond the individual level of analysis to look at variables that are expected to match at the organization and event levels of the Registry data.

Recommendation 9: Discussions should be held with the Registry office to find ways of reducing the complexity and time required to prepare data for analyses. To compare two or more variables from different analytic files, the files must be merged and restructured. Consideration should be given to the feasibility of Registry staff to perform some of the data restructuring prior to sending the export file to the data user, and if possible, to create export files already structured for variable matching.

Recommendation 10: Conduct continuous monitoring of data integrity, for example, by comparing verified with self-report variables, or the expected with the actual matches. It should be possible to model the difference between verified and self-reported variables. That is, use the known lag in self-reported vs. verified variables of the same kind and make use of this knowledge to determine, at any given point in time, expected matches.

and build this into the calculation of match percentages, to give a more realistic estimate of whether the actual data are consistent with what would be expected.

Recommendation 11: Expand the comparisons of data for internal consistency at the different levels of analysis (individual, organization, event). Identify variables and clarify assumptions regarding the degree of matching.

VI. Overall Conclusions

Overall, the results of this DQA reveal good to excellent integrity and completeness, lower levels for validity, and feasibility seems low, as currently viewed through the lens of a researcher. These findings indicate a robust database with more than adequate levels of data quality that, with greater use by researchers and analysts, is ready for continual improvement over time.

The results of the DQA highlight an important issue; the distinction between an administrative database as opposed to one closer to a public use dataset designed for third-party use. The Registry was designed to function as an administrative dataset, and this determined how fields and descriptors captured the data, with the tradeoff involving somewhat less structure, particularly for data users seeking to use the data for analytic or other purposes. But this purpose may not be completely consistent with the utility of a dataset in which data users will conduct statistical analyses. For example, the native format of data stored in the Registry, primarily text or string formats, is not optimal for data users, who will likely import a requested dataset into statistical analysis software. They will want data that consist of numeric formats, especially for single response labels, or formats appropriate to the analytic utility of the variables, such as dates or currency. Data users will also want multiple response variables to be clearly formatted as binary-coded individual variables. This will enable each of the multiple categories to be analyzed by the total number of responses rather than by the number of cases. Another example has to do with how missing data are handled. Whereas, for most administrative datasets, data that are missing show as blanks in the data file, researchers will want to classify missing data according to the reason they are missing, to distinguish valid missing data from data missing due to fields accidentally skipped, refusal, or "don't know" responses. An administrative dataset often does not provide enough information to give data users on why a given variable was blank for a particular case in the database.

In contrast to how an administrative dataset is used, a public use dataset has clearly specified variable formats, labels and a data structure to facilitate data analysis. This is not to say that the Registry data should be structured as a public use dataset, which is usually constructed from a large-scale research study analytic dataset. The current structure of the administrative dataset can be maintained while making changes developmentally that can assist data users. At present, data users will need to conduct much more data manipulation and file conversion, making their actual use of the dataset more time-consuming and can, in the long-run, serve as a disincentive for researchers and analysts to request data from the Registry.

The results of this DQA can serve to initiate a conversation with F5CA, F5LA and Registry staff regarding ways to improve both data quality and ease of use for data users to request and obtain an exported dataset for research, evaluation and policy analyses. Ultimately, making use of a robust and complex dataset such as the Registry will be critical for the state's ability to monitor and evaluate state early care and education workforce investments. By creating a standard Technical Manual and Codebook, data users can gain an understanding of the Registry and, by focusing on improving data quality, they can have confidence in the results of policy and evaluation studies that employ these data. In turn, this will encourage further efforts to integrate early childhood data systems across the state. Investments in verification, data structures and integration of "big data" may have a variety of benefits, including improving the quality of the ECE workforce and identifying evidence-based training and professional development programs. Ultimately, analyses of Registry data can contribute towards a qualified, well-prepared early care and education workforce.

Limitations. There are several limitations to this DQA that prevent drawing definitive conclusions. The most important limitation is that it was not clear whether the data delivered in the form of different Excel export files comprised all cases for the given dataset at the individual, organizational or event levels. This issue is known as the comprehensiveness of the records included in a given export file. All records are present and none are held back or filtered out. For example, when looking at the number of Registry participants with degrees it is important to consider whether or not the total numbers for each of the degree categories are what would be expected based on the number of individuals with a Registry profile. One potential indicator that not all cases were included occurred in the slight increase of cases as individual-level files were merged, but there could be a variety of reasons for this effect. Overall, this DQA was not able to determine the comprehensiveness of the data delivered for this analysis and was beyond the scope of this DQA. In the future, data users should be given information about the comprehensiveness of the data requested.

Comprehensiveness may also be due to the time lag between when verified vs. self-reported cases appear in the datafile. It is still possible that any differences between self-report and verified data may be due to this time lag. The same may occur where the export files included only cases with complete profiles, as opposed to partial profiles of Registry participants. In the future, it would be important to include specific information with every export file resulting from a data request, identifying whether the file contains all cases or, if requested by the data user, which cases were included and whether any filtering out of cases was done.

The challenge is not only whether or not all cases are included in a given export file, but rather whether or not a data user request, and the resulting export files that are sent, contain information on which cases were included, whether the data user requested current or active cases as opposed to inactive cases, and any other filtering that may have been requested. But, in fairness, no procedures were developed for the Registry office to provide this information to the study team when delivering the requested export files. This issue reflects a number of lessons learned in conducting the DQA. Other lessons include the following: The analysis plan for the DQA did not sufficiently foresee some of the challenges, such as the need to format and re-structure the Registry database and variable attributes, handling of multiple response variables, and the time and effort required for file merging and restructuring. Finally, we did no use the revised data request form, since it was still under development, in order to test feasibility. As a result, we may not have provided sufficient guidance to the Registry office both around the data request and their fulfillment of the request.

Finally, it is important to be careful interpreting the matching analysis for internal consistency for a number of reasons. First, only four pairs of variables across two content areas were compared and only at the individual level, so we cannot make any conclusions for other variables that might be compared, at this level and at other levels of analysis (organizational and event). Second, there could be valid and varying reasons for match percentages that are relatively low and may not indicate a potential lack of consistency. These analyses cannot identify the reasons for a lack of match particularly since, especially in the case of languages, there were many languages listed by users. Whenever a large number of categories are compared, the probability of finding a match will decrease, which may explain the lower match percentages for the language variables.

There is another challenge that is not directly related to the DQA but rather has important implications for interpreting the results from analyses of Registry data. As identified earlier in the Technical Manual, Registry data may not necessarily be representative of the entire CA ECE workforce and thus any results arising from analyses of Registry data cannot be validly generalized to all workforce members in the entire state. This issue is primarily due to the voluntary nature of enrollment in the Registry, which provides only a selected view of the ECE workforce across state. Additionally, this selectivity makes it difficult to obtain county-level estimates, which requires sufficient sample sizes at the county level. However, with enrollment increases, and if/when participation is made mandatory, it may be possible in the future to make claims about the generalizability of the results to the workforce as a whole across the state, and even, farther in the future, it may be possible to provide estimates for individual counties. As more organizations become part of the Registry then the representativeness of the sample at each level will improve.

VII. List of Recommendations

This section summarizes all of the proposed recommendations from this data quality assessment. Some of these recommendations might be accomplished in the short-term but, for others, they may require more time before any changes can be implemented. The short-term refers to approximately a one-year period for which recommendations can be addressed. However, the longer-term refers beyond the one-year period because these may require additional discussion, planning and decision-making, taking into account the current priorities and capacity of the Registry office, before any changes can be implemented.

Feasibility

Recommendation 1: All variable names in the export files should be checked against those listed in the Codebook. For example the variable names used in the export file should be checked against those listed in the Codebook so that there are identical names and formatting consistent with that in the Codebook. Systems and processes need to be in place to ensure the consistency is maintained as the Registry database is updated.

Recommendation 2: The Registry office should provide more information to describe export data when fulfilling a data request. For example, when an export file is delivered to a data user, it should contain a "transmittal form" identifying criteria used to include or exclude cases. It would be important to know whether the export files contain all current/active cases or if there were other characteristics that distinguishes the cases in the export files sent to data users.

Recommendation 3: The Registry database should be restructured to handle multiple response variables, in order to reduce file preparation in Excel and SPSS prior to statistical analyses.

Completeness

Recommendation 4: Reduce the amount of blank fields to increase information value. For variables with skip patterns faced by users for whom the variable does not apply, rather than leave these blank, unique values should be assigned as "Not Applicable" so that these can be distinguished from fields that are left blank, either by mistake or due to refusing to provide the information. As well, all blanks should be examined for other forms of missing values, such as "Don't Know" or "Don't have this information" as appropriate and assigned unique values for all variables, so that blanks would remain only for those fields that are accidentally left blank by the user.

Recommendation 5: Procedures should be in place to monitor levels of missing cases and make corrections as needed so that whenever a data request comes in, there is confidence that the export files contain complete data.

Recommendation 6: Registry staff should create procedures for the ongoing monitoring of data comprehensiveness. Variables included in export files should represent all current and active cases, unless otherwise specified by the data user, and information given about whether the export files reflect only complete and active profiles or not, with clear definitions for complete data.

Validity

Recommendation 7: To reduce the possibility of out-of-range values, the use of free-form fields for users to enter information should be curtailed so that the user interface offers fixed responses and formats for the user to enter. Also, procedures should be in place to ensure automatic checking as data are entered to fix out-of-range values as they occur.

Recommendation 8: Response choices that allow users not to answer the question, such as "Do not know" or "Do not wish to voluntarily report" should be organized for all variables. For example, non-valid responses should be grouped at the end of the list of valid response choices for all variables, allowing data users to identify and exclude these values for most analyses.

Integrity

Recommendation 9: Discussions should be held with the Registry office to find ways of reducing the complexity and time required to prepare data for analyses. To compare two or more variables from different analytic files, the files must be merged and restructured. Consideration should be given to the feasibility of Registry staff to perform some of the data restructuring prior to sending the export file to the data user, and if possible, to create export files already structured for variable matching.

Recommendation 10: Conduct continuous monitoring of data integrity, for example, by comparing verified with self-report variables, or the expected with the actual matches. It should be possible to model the difference between verified and self-reported variables. That is, use the known lag in self-reported vs. verified variables of the same kind and make use of this knowledge to determine, at any given point in time, expected matches.

and build this into the calculation of match percentages, to give a more realistic estimate of whether the actual data are consistent with what would be expected.

Recommendation 11: Expand the comparisons of data for internal consistency at the different levels of analysis (individual, organization, event). Identify variables and clarify assumptions regarding the degree of matching.