Inter-Rater Reliability Analysis of Data to Document the Consultative Examination Process

Volume 1: Final Report

November 4, 2012

David Wittenburg, Ph.D.
Debra Wright, Ph.D.
Sloane Frost, M.P.P.
Gordon Steinagle, D.O.
Ron Fine, M.D.

This page intentionally left blank for double-sided copying.

Contract Number:
SS00-09-60106

Mathematica Reference Number:
06691.900

Submitted to:
Social Security Administration
Office of Program Development and Research
6401 Security Blvd.
Robert M. Ball Building
Baltimore, MD 21235
Project Officer: Thomas V. Rush, Ph.D.

Prime Contractor:
Comprehensive Occupational Medical Services
51 Webster St
North Tonawanda, NY 14120

Submitted by:
Mathematica Policy Research
P.O. Box 2393
Princeton, NJ 08543-2393
Telephone: (609) 799-3535
Facsimile: (609) 799-0005
Project Director: David Wittenburg, Ph.D.

Inter-Rater Reliability Analysis of Data to Document the Consultative Examination Process

Volume 1: Final Report

November 4, 2012

David Wittenburg, Ph.D.
Debra Wright, Ph.D.
Sloane Frost, M.P.P.
Gordon Steinagle, D.O.
Ron Fine, M.D.

This page intentionally left blank for double-sided copying.

ACKNOWLEDGEMENTS

This report reflects the combined efforts of several people at the Social Security Administration, Comprehensive Occupational Medical Services, and Mathematica Policy Research, as well as several independent consultants. The study involved a large data extraction effort of sensitive information that had to be carefully planned and coordinated by management, research, clinical, and other staff. We are grateful to everyone for their efforts, though we might not get to name everyone individually below.

Of particular note was the guidance from our project officer, Thomas Rush, who provided important operational and technical guidance throughout the project, including leading several coordination activities within SSA to make the study possible. His comments on the data instruments, suggestions for describing the policy environment, and detailed comments on drafts of this report have considerably strengthened the evaluation findings.

The development of the data collection instruments, review of materials, and findings reflect the combined contributions of several staff in multiple offices at SSA. The project was led from the Office of Program Development and Research. It was initially developed by Richard Balkus. Michael Anzick oversaw the management activities throughout the project and provided important timely feedback on contractual issues needed to address the complex data extraction effort. Renee Ferguson, from the Office of Program Development and Research, provided valuable technical assistance throughout the project and led the sampling effort for the project. Deborah Harkin, also in the Office of Program Development and Research, provided disability examiner perspective guidance and reviewed folders as part of the data extraction for this report. Several SSA staff provided comments on this report that helped improved the substance. These reviewers included Rosemary Hall from the Office of Disability Programs, William Powell from the Office of Disability Adjudication and Review and staff from other offices. Finally, we are grateful to staff at the Office of Medical and Vocational Expertise (OMVE) who assisted in coordinating the activities for the SSA medical consultants. Robert Emrich, Jr., John Delpaine, and Gary Rauch provided management support throughout the project to ensure that the resources would be necessary within OMVE to conduct training and for medical consultants to review CEs. Diane Biller, Barbara Gall and Shelby Fizer provided assistance in coordinating times for the SSA medical consultants to extract data and provided technical support throughout the project.

At Mathematica, Claudia Gentile provided extremely useful quality assurance comments on a draft of this report. Mark Brinkley, Ryan Jackson, and Nadia Goranova led the development and revision of the website for the data collection. They each provided timely feedback to ensure that the website operated efficiently and securely to meet the needs of the project, especially by answering questions while the website underwent significant revision. Finally, Mark Beardsley was instrumental in conducting the programming for the entire study, and was especially helpful in producing statistics to assist our team in identifying problems with earlier versions of the instruments.

At COMS, we thank Michelle Dickerson and Lisa Skalman for their efforts in coordinating efforts for COMS's medical consultants. Michelle and Lisa diligently tracked all of the data entries from the pre-test and were always a cheerful voice in scheduling times for meetings and conference calls.

Finally, we thank all of the medical consultants and examiners from the COMS and SSA review teams who provided the data that are the substance of this report. As demonstrated in this report,

these reviewers provided expert advice in developing the data instrument and extracted data in a reliable manner. They provided the important information on CE processes, content, and quality that made this analysis possible.

This page intentionally left blank for double-sided copying.

CONTENTS

EXHIBITS

This page intentionally left blank for double-sided copying.

ACRONYMS

ALJ	Administrative Law Judge
ABMS	American Board of Medical Specialties
CE	Consultative Examination
CFR	Code of Federal Regulations
COMS	Comprehensive Occupational Medical Services
DDS	Disability Determination Service
Examiner	Disability Examiner
IOM	Institute of Medicine
IRR	Inter-rater reliability
MEDIB	Management Information eDIB
MER	Medical evidence of record
MSS	Medical Source Statement
ODAR	Office of Disability Adjudication and Review
OMVE	Office of Medical and Vocational Expertise
OPDR	Office of Program Development and Research
POMS	Program Operations Manual System
SSA	Social Security Administration
SSDI	Social Security Disability Insurance
SSI	Supplemental Security Income
SLR	Straight leg raise

This page intentionally left blank for double-sided copying.

I. INTRODUCTION

The Social Security Administration (SSA) administers the Social Security Disability Insurance (SSDI) and Supplemental Security Income (SSI) programs. The SSDI program pays disability benefits based on a person's work history, whereas SSI disability benefits are paid to people with limited income and resources. The disability eligibility requirements for adults require that an individual be unable to "engage in any substantial gainful activity by reason of any medically determinable physical or mental impairment which can be expected to result in death or which has lasted or can be expected to last for a continuous period of not less than 12 months" (Section 223 (d)(1)(a) Social Security Act).[1] The SSA disability determination process includes an initial disability determination and a procedure for appeals which can occur at the initial level or a hearings level.

The consultative examination (CE), which is a physical or mental health examination or test purchased on behalf of the claimant at SSA's expense, is an important component of the disability determination process. It is requested by the agency to make a disability determination. The state Disability Determination Services (DDS), which are federally funded state agencies that support SSA in making disability determinations, manage the process of ordering and paying for CEs. The claimant can personally supply medical evidence of record (MER) to the DDS, but the DDS usually obtains medical evidence directly from medical sources, which can be the claimant's treating source or another medical provider. The DDS will order a CE from a CE provider, if it is necessary to obtain additional information to make an informed disability determination. SSA and its state DDS agencies ordered CEs for approximately 48 percent of 2009 disability claims, which represented over one million CEs (Social Security Advisory Board 2012, table 46).

The Code of Federal Regulations (CFR) describes the federal regulatory guidelines governing CEs processes and content for the SSDI and SSI programs in §404.1519 and §416.919, respectively.[2] SSA is expected to oversee and assess the CE process and content in accordance with these regulations, which are specified under Title 20 of the CFR §404.1519 and §416.919.[3]

SSA awarded a contract to Comprehensive Occupational Medical Services (COMS), which subcontracted with Mathematica Policy Research, to extract information from a sample of claims closed at two administrative levels (initial and hearings) in 2009 that contained CEs procured to assist in the determination or decision of the claim at that level. We refer to the data extracted as the *CE Review*. The CE Review covers three topic areas:[4]

---

[1] The programs also have separate definitions for other groups, including youth. We focus on the adult definition for the purpose of this report. For the official SSA operational definition of disability, see https://secure.ssa.gov/poms.nsf/lnx/0400115015 (accessed July 11, 2012).

[2] A link to the federal regulations is available at http://www.socialsecurity.gov/OP_Home/cfr20/416/416-0919.htm (SSDI) and http://www.socialsecurity.gov/OP_Home/cfr20/404/404-1519.htm (SSI) (accessed August 13, 2012).

[3] Specifically, according to the CFR §404.1519t and §416.919t, SSA needs to (1) ensure that referrals for and purchases of CEs are made in accordance with SSA requirements, (2) monitor both the referral processes and the product of the CEs obtained, and (3) perform ongoing special management studies of the quality of CEs purchased from key providers and other sources and the appropriateness of the examinations authorized.

[4] The original study questions developed by SSA for the study included 15 topic areas with 38 specific questions.

1. **CE processes:** Are CEs being requested in compliance with federal regulations?
2. **CE content:** Are CE providers including content in compliance with federal regulations?
3. **CE completeness and quality:** Do CEs include sufficient information to make a disability determination, and did SSA receive everything it paid for in the exam? Additionally, are there process and content factors that contribute to the quality of CEs?

A challenge in extracting information on CEs for the CE Review is that they are stored as scanned medical records in a disability claimant's permanent record. The content included in CEs might vary depending on the claimant's impairment. Additionally, the forms used to document CE processes and content might vary substantially, and most text, particularly for medical content, is in narrative format.

To address this challenge, our "design team" developed a set of instruments that could be used by "review teams" to extract information on CEs. The review teams included disability examiners ("examiners") and consultants with medical or psychological backgrounds ("medical consultants"). The study included separate review teams from COMS and SSA to assess the comparability of data extraction for a common sample of CEs. With input from SSA staff, the design team developed a template of questions, which included "instruments" for medical consultants and examiners. The examiner instrument included questions on CE processes, such as type of CE and scheduling dates. The medical consultant instrument included questions on CE content and quality, including any items that required medical judgment. The review teams answered the questions using a secure "web-based template," which was a secure website that the review teams accessed by entering a login ID and password. Specifically, the web-based template included the examiner and medical consultant instruments, which medical consultants and examiners could access by logging into the website. The instruments for the web-based template underwent several rounds of review. We refer to the data extracted using the final instruments as the "CE Review data."

We compare the data extracted by the review teams in an inter-rater reliability (IRR) analysis to assess whether the responses from COMS's medical consultants and examiners align with those from SSA's medical consultants and examiners. A high level of agreement would indicate that the data were entered reliably, whereas a low rate would indicate that the data were measured with substantial variability.

The design team had to revise the instrument multiple times before achieving this strong rate of agreement, the data extraction process was cut short due to the end date of the contract. The initial data extraction effort started in Fall 2009 but was stopped after an analysis of a limited sample of CE Review data indicated a low level of agreement between COMS and SSA medical consultants. In response, there was a lengthy revision process that included focus groups to assess the problems, revisions to questions, multiple pretests, the development of a codebook, and a final training.

The fully revised final instruments were ready in Fall 2011. The final sample for the IRR analysis included 289 adult CEs matched pairs for the medical consultant instrument.[5] This sample

[5] While our team collected information on a small number of child CEs (11 CEs), we restrict our analysis to the large sample of adult CEs.

included CEs from the two largest physical health exam categories (internal medicine and musculoskeletal) and mental health exams from the initial and hearings level. During the development of the final instruments for the medical consultants, SSA requested a small scale IRR analysis of the examiner instrument to ensure it was operating as envisioned, which included 25 adult CEs matched pairs.[6]

This report, which is the first of two final reports in this study using the CE Review, summarizes findings from the IRR analysis of the examiner and medical consultant instruments. The primary objective of this report is to provide an overall assessment of the reliability of the data entry. The second report for the study will provide a descriptive summary of findings to address the three areas of research questions identified above related to CE processes, content, and completeness.

Our IRR findings indicate a strong rate of agreement for the final CE Review medical consultant and examiner instruments. Of the 142 questions in the final examiner and medical consultant instruments, 128 questions exceeded the reliability thresholds, which we generally defined as a percentage agreement of at least 70 percent (i.e., the data extracted from COMS medical consultants and examiners perfectly matched the data from SSA medical consultants and examiners for at least 70 percent of CEs).[7] These findings are impressive given that the review teams had to extract information from scanned medical records.

The findings have important implications for producing statistics using the CE Review data. The strong IRR findings provide an indication of the quality of the instrument's items and its utility as a data extraction tool. They indicate it is possible to extract reliable data, at least for the largest exam types and across adjudication levels. We do not have information on how the instruments would perform for other groups (e.g., pediatrics or smaller physical health groups). Nonetheless, the strong performance for the largest subgroups included in our CE sample suggests that the instruments developed for this study hold strong promise for future data extraction efforts.

To maintain the long-term objectives of the CE Review, SSA requested that the web-based template be transferred to SSA for potential use in future work. The transfer of the website to SSA is especially important given that only a limited sample of CEs was available for this study. SSA can use the findings and data from the IRR study in future data extraction efforts to check the content and quality of data collected by future review teams. For this reason, we provide a detailed analysis of the IRR findings for all examiner and medical consultant questions. These detailed findings provide an important source of information for future review teams, particularly if SSA is interested in changing the wording of any questions to meet their long-term research needs. We also include additional documentation that could be helpful to future review teams, including the codebook, in the appendices.

---

[6] An IRR analysis for the examiner instrument was not originally planned because the design team assumed that the examiner items were generally objective in nature and hence, should have very few reliability problems. However, when significant problems in the medical consultant instrument emerged early in the study, SSA requested a small scale test of the examiner instrument to ensure it was operating as envisioned.

[7] Of the 128 questions that were judged as meeting our reliability standards, 126 were above 70 percent agreement. We judged two additional variables that had continuous response options (e.g., dollar amounts) as reliable based on having a correlation coefficient above 0.9. As will be described in Chapter IV, we use percent agreement for most of the outcomes with dichotomous response options. However, we use correlation coefficients for a limited number of questions with continuous response options (e.g., dollars).

The report includes two volumes. Volume 1 is intended to provide readers a summary of the general findings from the data extraction including an overview of the CE process (Chapter II); the development of the data extraction instrument and sample frame (Chapter III); the methodology for conducting the IRR analysis (Chapter IV); a summary of IRR findings (Chapter V); detailed IRR findings (Chapter VI); and a summary discussion of the IRR analysis and future directions (Chapter VII). Volume 2 includes the medical consultant and examiner data instruments, including skip patterns for the questions (Appendix A); the codebook for the medical consultant instrument (Appendix B); and additional findings from the IRR analysis (Appendix C).

II. DISABILITY DETERMINATION PROCESS AND CONSULTATIVE EXAMINATIONS

This chapter provides an overview of the disability determination process, the role of the CE in this process, and the implications of these processes for our data extraction efforts. Our background description on CE processes and content is based primarily on the federal regulations and information from "Consultative Examinations: A Guide for Health Professionals," which is referred to at SSA as the *Green Book*.[8] The federal regulations and the Green Book text, which are cited in detail below, are both available online.[9] The federal regulations for CEs are described in §404.1519 (for SSDI) and §416.919 (for SSI). These regulations outline the general processes and content that should be included in a CE. Federal regulations do not specify the exact content of a CE report. The definitive source of this policy guidance is the Program Operations Manual System (POMS). The Green Book is a handbook used by the DDS to train new CE providers. The decision to use this handbook as the study source material was based on the reality that CE providers use this source (rather than the POMS) to prepare their reports.[10]

A. Overview of the Disability Determination Process

The SSA disability determination process includes an initial level of consideration and a procedure for appeals. Below, we discuss the application processes associated with the initial and hearings levels.

1. Initial Level Cases

The application process for SSDI and SSI usually starts with the submission of a claim to a local SSA field office, which we refer to as the *initial level*. Claimants may file claims in person, by telephone, by mail, or online. The claim includes information about the claimant's medical and nonmedical allegations necessary to determine benefit entitlement. The field offices are responsible for verifying the nonmedical eligibility requirements (such as age, employment, recency of earnings, and income) for either program.

After reviewing the claim, the field office sends the case to the appropriate state DDS for medical adjudication.[11] The DDS is federally funded, though the structure of these agencies varies.

---

[8] In developing the instruments for the web-based template, we also cross-checked items from the CFR and Green Book with the Program Operations Manual System (POMS), which provides internal guidance on SSA's procedures, and the Hearings, Appeal and Litigation Law (HALLEX), which includes procedures that Administrative Law Judges (ALJs) should use to order CEs. POMS is available at https://secure.ssa.gov/poms.nsf/home!readform, and HALLEX is available at (http://www.ssa.gov/OP_Home/hallex/I-02/I-2-5-20.html (accessed August 13, 2012).

[9] A link to the federal regulations is available at http://www.socialsecurity.gov/OP_Home/cfr20/416/416-0919.htm (SSDI program) and http://www.socialsecurity.gov/OP_Home/cfr20/404/404-1519.htm (SSI program) (accessed August 13, 2012). A link to the Green Book is available at http://www.ssa.gov/disability/professionals/greenbook/index.htm (accessed July 10, 2012).

[10] According to SSA staff, efforts are underway to update the Green Book.

[11] The names of the state agencies that administer this process vary from state to state. For example, Florida calls its agency the Division of Disability Determinations while New Mexico uses Disability Determination Services for its agency. For ease of exposition, we use the term DDS to describe all such state agencies. A detailed list of agencies is available at http://www.ssa.gov/disability/professionals/procontacts.htm (accessed July 1, 2012).

Some have a centralized system within one DDS office; others have a decentralized system of several offices within the state. Once the state agency makes a disability determination, it returns the case to the field office, which notifies the claimant of the outcome.

Within the DDS, examiners collect information used to make determinations on the claimant's medical eligibility. According to SSA staff, the average DDS office has 130 examiners, though there is substantial variation by office. The examiner is responsible for gathering the claimant's medical records and acquiring additional information, such as vocational information. Examiners face increasing pressure to process a large volume of cases because the number of SSDI and SSI claims has increased significantly over the past decade, particularly in recent years following the economic recession in 2008 (Social Security Advisory Board 2012).

The examiner usually works with a medical or psychological consultant, which we refer throughout the report as a medical consultant, to make a disability determination. Every DDS has a team of medical consultants and examiners who help determine whether claimants meet the disability eligibility criteria. According to an IOM (2006) report that focused on redesigning the disability determination process, in 2004, state DDS agencies had more than 2,100 medical consultants across the country. All DDS agencies had medical consultants in the clinical areas covered by most claims (e.g., mental health, internal medicine, and pediatrics), though the number of specialists outside these areas varied substantially. For example, IOM found in 2004 that 29 DDS agencies had no medical consultants specializing in cardiology, 28 had no neurologists, and 25 had no orthopedic surgeons or orthopedic specialists.

Examiners and medical consultants base their determinations on medical evidence. Claimants' records from their health care providers, referred to as *treating sources* by SSA, are usually considered the best source of medical evidence for the case. If SSA finds that a treating source's opinion on the issue(s) of the nature and severity of the claimant's impairment(s) is not inconsistent with substantial evidence in the record, SSA will give their opinions controlling weight.

In documenting the medical determination, the examiner should indicate whether and how the claimant's impairments satisfy the eligibility requirements. The examiner consults with the medical consultant on the nature and severity of the impairments as well as what kind of additional medical evidence is needed to decide the case. In general, the examiner should not make determinations on medical eligibility without consulting with the medical consultant.[12] There are exceptions when an examiner can act on his or her own, such as in quick disability determination or single decision maker cases.[13]

In most states, claimants who are denied at the initial level can appeal their determination and request reconsideration by the DDS.[14] For cases reconsidered by the DDS, claimants provide the

[12] For more information on the medical consultant's role, see the POMS's description at https://secure.ssa.gov/apps10/poms.nsf/lnx/0439518010 (accessed August 13, 2012).

[13] For more information on these cases, see http://www.ssa.gov/disabilityresearch/qdd.htm and http://www.ssa.gov/OP_Home/cfr20/404/404-0906.htm (accessed July 10, 2012).

[14] Ten states are participating in the disability redesign prototype model ("prototype DDS"), where they send initial determination appeals to the Office of Disability Adjudication and Review (ODAR) for a hearing (the hearings process is described below). For more details, see https://secure.ssa.gov/poms.nsf/lnx/0412015100. These ten states are also part of the larger group of twenty states described below that are testing disability process initiatives.

field office with any new information to support their cases. Such new information usually consists of documentation about receiving additional treatment or having seen an additional treating source. The field office sends the new information to the DDS for a second medical review by a different examiner-medical consultant team. If the claimant said that he or she received additional treatment, the DDS must attempt to obtain evidence of this treatment before making a determination.

In twenty states, SSA is testing disability process initiatives to improve the disability determination process, which can influence the order of a CE. [15]Some of those tests have been stand-alone tests, while others test various combinations of modifications to the disability determination procedures. Most notable for the ordering of the CE in these states was the modification for the "single decisionmaker model," which allows examiners more control of decisions to order CEs without signoff from a medical consultant. Specifically, in the single decisionmaker model, qualified examiners are given authority to complete all disability determination forms and make initial disability determinations in many cases without medical consultant signoff.

2. Hearings Level Cases

Hearings level appeals occur outside the DDS in SSA's ODAR. If a claim is denied by the DDS as a result of the reconsideration in a non-prototype DDS or the initial determination in a prototype DDS, the claimant may ask for a hearing before an Administrative Law Judge (ALJ). This is the first point at which claimants, in addition to submitting medical records for the case, may appear in person to discuss specific elements of their medical history for the disability determination. ALJs conduct hearings and render case decisions. At the hearing, claimants and their representatives may present new medical evidence to support their cases. For cases appealed to a hearing, expertise can come from either or both medical or vocational experts who agree to testify as expert witnesses. The ALJ makes a decision concerning the case and notifies the claimant in writing. Those denied by the ALJ can appeal to the Appeals Council, which acts as the final level for review within SSA. If the Appeals Council decides to review the case, it will either decide the case itself or return it to an ALJ for further review.

There is an additional stage of appeals in the federal court system.[16] If a claimant disagrees with the Appeals Council's decision or if the Appeals Council decides not to review the claimant's case, the claimant may file a lawsuit in federal court and pursue that case through all appeals levels.

The number of cases is substantially larger at the initial level than at the hearings level. In 2009, there were 3 million initial level cases in comparison to 652,000 hearings level cases (SSA 2011 Titles II and XVI Disability Research Files).[17] This large difference between levels is important context for our findings, because our sample (described in Chapter III) includes an oversampling of hearings level cases.

---

[15]For more details, see https://secure.ssa.gov/poms.nsf/lnx/0412015100.

[16] For more details, see http://www.ssa.gov/pubs/10041.html#a0=1 (access August 13, 2012).

[17] SSA provided these statistics in response to a request made for this project. Additional citations for application statistics are available in Social Security Advisory Board (2012).

B. Role of Consultative Examinations in Disability Determinations

The DDS must follow general SSA operational guidelines in ordering CEs, though it has flexibility in the forms used to document the process and in maintaining relationships with CE providers. Below, we describe these processes and then discuss their implications for the development of the examiner and medical consultant instruments.

1. DDS Processes in Ordering CE

The rules and procedures for requesting a CE apply to both the initial and hearings level, and federal regulations specify the conditions under which a CE should and should not be ordered (§404.1519a, §404.1519b, §416.919a, and §416.919b). The decision to order a CE is based on full consideration of whether additional information (such as clinical findings, laboratory tests, diagnosis, and prognosis) is needed to adjudicate the case.

At the initial level, the decision to order a CE involves an examiner, often working with a medical consultant at the DDS to determine if a CE is necessary. In general, if an examiner determines that a claimant does not have an acceptable medical source or that the medical evidence contained in the case folder is insufficient to make a medical determination, the examiner will schedule a CE.[18] As with other elements of the disability determination process described above, the examiner should consult with a medical consultant before gathering medical evidence, including ordering a CE, to ensure that the consultant agrees with the decision.

The decision to order a CE at the hearings level is made by the ALJ, though the CE is ordered through the DDS. According to the Green Book, medical development at the hearings level frequently is conducted through the DDS. However, hearings offices may also contact treating sources directly. In rare circumstances, an ALJ may issue a subpoena requiring a claimant to produce evidence or provide testimony at a hearing.

The CE order should only request information to adjudicate the case and avoid ordering any unnecessary or invasive procedures (§404.1519f and §416.919f). This requirement has important implications for the analysis of CE quality, because it implies that the amount of information needed varies by CE. For example, the DDS might order a CE to obtain one or more ancillary studies, which likely would be substantially shorter than a CE for a full internal medicine exam.

DDS examiners document the process and rationale for ordering the CE using a worksheet format that varies by state. There are no operational requirements available to us in the electronic case record (eView) to document the CE request or to provide a rationale as to why one is being ordered. For example, some states require supervisory approval for new examiners when ordering a CE; an experienced examiner, however, may order CEs without such approval. Worksheet formats

[18] There are five general situations in which a CE might be needed: (1) the additional evidence needed is not contained in the MER; (2) the evidence may not have been available from the treating or other medical sources for reasons beyond the claimant's control; (3) highly technical or specialized medical evidence was not available from the treating source or other medical sources; (4) the MER has a conflict, inconsistency, ambiguity, or insufficiency that must be resolved; or (5) there was an indication of a change in the claimant's medical status, and the current severity level is not established.

vary substantially from narrowly defined ones that include specified CE-related data items (such as fees for exam) to completely open-ended ones.[19]

Some DDS agencies also provide standardized forms to CE providers to collect information on claimants, though the actual form and the information collected varies by state. For example, Maryland created its own standard form for assessing joint range of motion that is sent to CE providers when this information is relevant to making a disability determination.

2. CE Provider Qualifications

Once a decision is made to order a CE, whether at the initial or hearings level, DDS examiners coordinate the process by sending relevant information to a CE provider who is a "qualified medical source" (§404.1519g and §416.919g), which can either be the claimant's treating source (the preferred choice) or a state-contracted CE provider. Regulations define qualified medical sources as professionals who are licensed in the state and have the training and experience to perform the examination or test requested. When possible, the federal regulations note there is a preference to receive the CE from the claimant's treating source (§404.1519h and §416.919h), though they also specify the conditions under which other medical sources can be used (§404.1519i and §416.919i).[20] The Green Book notes that medical sources are selected based on appointment availability, distance from a claimant's home, and ability to perform specific examinations and tests for the fee specified in the state fee schedule.[21]

The DDS has flexibility in managing CE programs with providers but must maintain minimum standards in managing this process, and the providers must meet minimum qualification standards (§404.1519s and §416.919s). Each state is responsible for comprehensive oversight of its CE program, with special emphasis on eligible providers. The DDS must maintain an active process for recruiting CE providers, ensure that these providers are appropriately trained, and monitor providers to ensure that the CEs are performed in accordance with regulations.[22] Eligible providers must meet

[19] At the time of our review, the two most common worksheet formats were VERSA and LEVY. The VERSA worksheet included several specific CE-related data items (such as CE fees), whereas the LEVY worksheet was a largely blank page. Several states, such as New York and California, have their own forms. In July 2009, the DDS legacy system for worksheets was updated. The VERSA is now known as Iron Data-Toronto, and the LEVY is now known as Iron Data-St. Louis.

[20] The regulations note that another medical source might be used in four cases: if (1) the claimant's treating source prefers not to perform such an examination or does not have the equipment to provide the specific data needed; (2) there are conflicts or inconsistencies in the case that cannot be resolved by going back to the treating source; (3) the claimant prefers a source other than the treating source and had good reason for the preference, or (4) prior experience indicates that the treating source may not be a productive source (for example, if he or she has consistently failed to provide complete or timely reports) (§404.1519i and §416.919i). Under certain conditions, a claimant can raise objections to the medical source selected as the CE provider (§404.1519j and §416.919j).

[21] For details, see this section of the Green Book: http://www.ssa.gov/disability/professionals/greenbook/ce-guidelines.htm (accessed August 14, 2012).

[22] Specifically, federal regulations require the DDS to maintain "(1) An ongoing active recruitment program for consultative examination providers; (2) A process for orientation, training, and review of new consultative examination providers, with respect to SSA's program requirements involving consultative examination report content and not with respect to medical techniques; (3) Procedures for control of scheduling consultative examinations; (4) Procedures to ensure that close attention is given to specific evaluation issues involved in each case; (5) Procedures to ensure that only required examinations and tests are authorized in accordance with the standards set forth in this subpart; (6) Procedures for providing medical or supervisory approval for the authorization or purchase of consultative examinations and for

federal regulations. Several states work with volume providers who perform large numbers of CEs.[23] According to IOM (2006), these regulatory requirements are relatively minimal outside of being licensed in the state and having the training and experience to perform the type of examination or test being requested. In general, CE providers must have the facilities and equipment needed to perform the requested examinations or tests and have a good understanding of SSA's disability programs and their evidentiary requirements. However, they are not required to have specific formal training or certification in the evaluation of disability.

The final step in the CE ordering process is for every CE report to be completed and properly signed (§404.1519p and §416.919p). If the CE is inadequate or incomplete, the DDS will contact the medical source and ask him or her to furnish the missing information or prepare a revised report.

3. CE Content: Medical History, Exams, and Additional Tests

To develop a CE, the DDS sends CE providers background information on the claimant, including medical records submitted by the claimant as part of the claim (referred to as the *medical evidence of record* [MER]). The DDS provides background information to the CE source based on the specific case facts.

The CE provider should review any medical information sent from the DDS and verify the background information for the claimant. Specifically, CE providers should document the claimant's identification and note in their CEs whether the DDS sent any additional MER. The medical records provide important context that should facilitate an efficient examination, and they are important in documenting whether the claimant's complaints or exam results deviate from the MER.

The Green Book outlines the general medical content for the CEs, which should include the following:

- **Medical history.** The CE provider should document who provided the medical history (for example, the claimant or a third party) and provide an assessment of its reliability. CEs should include information about the claimant's present and past medical history, including the major or chief complaint; any other complaints; and additional history that might be pertinent to the claim (for example, prescription drug, family history, or drug or alcohol use).

(continued)
additional tests or studies requested by consulting medical sources; (7) Procedures for the ongoing review of consultative examination results to ensure compliance with written guidelines; (8) Procedures to encourage active participation by physicians and psychologists in the consultative examination oversight program; (9) Procedures for handling complaints; (10) Procedures for evaluating claimant reactions to key providers; and (11) A program of systematic, onsite reviews of key providers that will include annual onsite reviews of such providers when claimants are present for examinations." See http://ssa.gov/OP_Home/cfr20/416/416-0919s.htm and http://ssa.gov/OP_Home/cfr20/404/404-1519s.htm for more details.

[23] The federal regulations define an eligible volume provider (called *key providers* in the regulations) as meeting one of the following conditions: (1) a CE provider with an estimated annual billing to SSA disability programs of at least $150,000; or (2) a CE provider with a practice directed primarily towards evaluation examinations rather than the treatment of patients; or (3) a CE provider that does not meet the above criteria but is one of the top five CE providers in the state by dollar volume, as evidenced by prior year data. See http://ssa.gov/OP_Home/cfr20/416/416-0919s.htm and http://ssa.gov/OP_Home/cfr20/404/404-1519s.htm.

- **Physical or mental health exam findings (specialty exams).** There are specific CE guidelines in the Green Book for several different specialty exams, including internal medicine, rheumatology, musculoskeletal (orthopedic), respiratory, cardiovascular, and neurological exams, and for mental health exams.[24] Federal regulations (§404.1519n and §416.919n) also provide general guidelines on the length of the exam to ensure that such examinations are complete.

- **Additional tests (laboratory, X-rays, and psychological tests).** The CE should summarize the results of any laboratory and other tests (such as X-rays or psychological tests) ordered by the DDS. For example, a CE might require a detailed medical exam if the claimant's claim is missing detailed medical information. Conversely, a special test such as an X-ray, blood studies, or an electrocardiogram might be necessary to make a determination. According to federal regulations (§404.1519m and §416.919m), a CE should only include tests needed to make a determination, which vary by case. When ordering specific exams, the DDS must carefully consider the invasiveness or risk of the exam to the claimant.

- **Medical source statement.** CEs should generally include a statement about the activities the claimant can still perform despite his or her impairment(s), unless the case is based on statutory blindness. This is referred to as the *medical source statement* (MSS). According to the Green Book, the MSS should provide an opinion about the claimant's ability, despite his or her impairment(s), "to do work-related activities such as sitting, standing, walking, lifting, carrying, handling objects, hearing, speaking, and traveling; and, in cases of mental impairment(s), the opinion of the medical source about the individual's ability to understand, to carry out and remember instructions, and to respond appropriately to supervision, coworkers, and work pressures in a work setting."[25] The provider might use a multipage form that allows a physician to address a claimant's functionality and limitations for the MSS or submit it in another format, such as a letter. These forms are not standardized unless the source happens to obtain a copy of SSA's standardized MSS forms created by ODAR to standardized requests (HA-1151 and HA-1152). According to SSA staff, these forms are not actively promoted at the initial level.[26] Federal regulations (§404.1519n and §416.919n) also note that the SSA (DDS) will ordinarily request an MSS as part of the CE, though the absence of such a statement does not make the report incomplete.

[24] The detailed guidelines for adult CE exams are available online at http://www.ssa.gov/disability/professionals/greenbook/ce-adult.htm (accessed August 13, 2012).

[25] See http://www.ssa.gov/disability/professionals/greenbook/ce-adult.htm (accessed August 15, 2012).

[26] For details on the residual functional capacity forms, see https://secure.ssa.gov/poms.nsf/lnx/0480850025 (accessed September 10, 2012).

4. CE Completeness and Quality: Elements of a Complete CE Report

According to regulations (§404.1519n and §416.919n), a complete CE should, with few exceptions, generally include the following elements, most of which are described in Section II.C:[27]

- Major or chief complaint(s)
- **Chief complaint history.** A detailed description, within the area of specialty of the examination, of the history of the major complaint(s)
- **Discussion of findings.** A description and disposition of pertinent positive and negative detailed findings based on the claimant's history, examination, and laboratory tests related to the major complaint(s) and any other abnormalities or lack thereof reported or found during the CE or laboratory testing
- **Lab tests.** Results of laboratory and other tests (for examples, X-rays) performed according to the requirements stated in the Listing of Impairments[28]
- **Diagnosis and prognosis.** The diagnosis and prognosis for the impairment(s)
- **MSS** (optional)**.** A statement about what the claimant can do despite his or her impairment(s), unless the claim is based on statutory blindness (as noted in the previous section, however, the absence of such a statement does not make the CE report incomplete)
- **Signed report.** The report must be signed by the CE provider

5. Final Review of CE by DDS agencies

Once a provider completes a CE, the DDS is responsible for reviewing it for completeness and ensuring that it was signed. If any information is missing, the DDS needs to follow up with the provider to obtain the additional information or signature or to clarify a finding.[29]

C. Implications for Data Extraction

Variations in CE processes across states presented challenges for data extraction. Many CE records were in different formats, depending on the individual's type of exam and state of residence. The medical consultants also had to capture information about specific types of CEs, specialty exams, and lab tests that might affect CE content and quality. The lack of structure in the CE process created some initial challenges in designing a reliable instrument. As will be described in Chapter III, the design team eventually overcame these challenges by developing more specific

---

[27] These items directly correspond to the text in §404.1519n and §416.919n. However, we added short bullets to the summary of items included in the federal regulations to correspond with their appearance in our summary exhibit in Chapter VI.

[28] For details on the adult Listing of Impairments, see http://www.ssa.gov/disability/professionals/bluebook/AdultListings.htm (accessed August 17, 2012).

[29] For signature requirements, see http://www.socialsecurity.gov/OP_Home/cfr20/416/416-0919n.htm and http://www.socialsecurity.gov/OP_Home/cfr20/404/404-1519n.htm.

wording for questions and guidelines, which included a codebook to provide guidance on how medical consultants should extract data and a training session to address any questions.

This page intentionally left blank for double-sided copying.

III. INSTRUMENT DEVELOPMENT AND SAMPLE SELECTION

This chapter provides background information on the development of the instruments and the sample selection for this study. We begin by providing an overview of the initial study design and subsequent changes that the design team made to the instrument. We then review the data extraction process, which had to be substantially revised to include additional materials, such as a codebook, to facilitate data entry. The final data extraction also included a smaller sample than originally anticipated for the IRR analysis due to delays in developing the final instruments. We conclude with a summary of the final web-based template, which includes separate data extraction instruments for the examiner and medical consultant.

A. Study Objectives and Background

The design team developed separate instruments for medical consultants and examiners to address the research questions identified in the study. The initial sample targets included 6,000 CEs. To facilitate the data extraction process, the review teams from COMS and SSA included medical consultants and examiners with extensive practical knowledge of SSA's disability processes. Below, we provide additional details on the instruments, initial sample targets and review teams.

1. Instruments Set Up to Address Key SSA Questions

The SSA research questions for the project addressed some issues that applied to all CEs and others that applied to subgroups of CEs, such as CE type, adjudication level (initial versus hearing), or state of DDS. The number of questions in the instrument varied depending on CE type and the individual characteristics of the case, such as exams relating to particular health conditions. As outlined in Chapter I, the questions generally covered topics related to CE processes, content and completeness. Within these topics, the study had several specific sub areas of interest.[30]

To maximize the medical consultants and examiners' efficiency and reliability in addressing SSA's questions, the design team organized the instrument to reflect the way in which medical consultants and examiners would analyze and enter information for a claimant.[31] For examiners, the instrument included a relatively short battery of questions on the process of ordering CEs. For medical consultants, the instrument included much more detailed assessments on the content and quality of the CE, especially on the medical history, components of the physical or mental exam, and any additional tests.

2. Initial Sampling Goals for the CE Review Included 6,000 CEs

The initial plan for the study was to extract data for a cohort of CEs in claims with determinations and decisions closed at the initial and hearings level in 2009. The year of the

[30] As shown in Appendix Exhibit A1, the original study questions were divided initially into 15 areas with 38 specific questions. For simplicity, we group these questions into the three areas CE processes, content, and completeness.

[31] The initial data extraction instrument was designed for the medical consultant only and included 18 general sections that had alphabetic labels from A through Q. However, during the development of the instrument, a separate instrument was added for the examiner, which included some questions related to CE processes.

determination (for initial level CEs) and decision (for hearings level CEs) was the basis for inclusion. The initial plan was to split the cohort to include a review of 600 CEs as part of the IRR study and then review an additional 5,400 CEs for the full study (that is, a total sample of 6,000 CEs). The total sample was designed to produce representative statistics on CEs at the state level and include both initial and hearings levels.

3. COMS and SSA Medical Consultants and Examiners had Extensive Practical Experience with SSA Disability Determination Process

A unique feature of the data extraction effort was that both COMS and SSA review teams included medical consultants and examiners who had extensive experience in the disability determination process. The SSA medical consultants included staff from the Office of Medical and Vocational Expertise (OMVE) and an examiner from the Office of Program Development and Research (OPDR). The OMVE medical consultants, whose main responsibilities are to provide expert advice and support SSA's disability determinations process, had extensive experience in reviewing disability cases.[32] The COMS medical consultants had work experiences very similar to those of the SSA medical consultants.[33] Several COMS medical consultants also had practical experience in extracting information on CEs based on a previous study for SSA.[34]

B. Data Extraction Process

The sample for the study was selected from an internal SSA database called the Management Information eDIB (MEDIB) by a statistician and the review teams extracted information on the selected sample using a system called "eView". The MEDIB contains information on CEs and other

[32] OMVE manages a nationwide network of medical, psychological, and vocational experts who assist federal reviewing officials, ALJs, the Decision Review Board, state DDS agencies, and the Office of Quality Performance in making disability determinations and decisions. For more details on OMVE, see http://www.ssa.gov/org/orgdcrdp.htm#omve (accessed July 17, 2012). In total, 11 medical consultants participated in the medical consultant review of the 289 CEs covered in this report. A single OPDR examiner, who had previous experience in reviewing disability cases at the DDS level and OMVE, reviewed the 25 CEs included in the examiner review.

[33] All nine COMS medical consultants who participated in the medical consultant review of 289 CEs covered in this report had extensive experience reviewing CEs, and most had direct experience reviewing CEs while working at OMVE or at a DDS. The one exception had extensive experience reviewing disability claims in other capacities, including as a medical consultant in a previous data extraction effort led by COMS (described above). The COMS examiner review team included three people who had to review a larger sample of CEs to identify their characteristics for the larger medical consultant study. The COMS examiners included two examiners from the Maryland DDS who reviewed all CEs, except those from Maryland. A COMS medical consultant reviewed the Maryland CEs.

[34] Under a previous contract completed in April 2008, COMS medical consultants completed a data extraction effort and evaluation of the CE processes and content using a sample of approximately 1,500 CEs from administrative folders at the initial level. The design of the data extraction effort in this contract differs in five important ways from the previous effort. First, the design team developed questions for the instrument in conjunction with SSA to address a broad set of research questions that were not covered in the first report. Second, SSA based the sample selection criteria on the new electronic folder system to ensure a random selection of CEs. Third, the sample included initial and hearings level CEs. Fourth, the instrument included separate sections for examiner and medical consultants. The examiner's role was introduced to make data extraction more efficient. As will be discussed in more detail below, the examiner collected information to identify the type of CE that was eventually shared with the medical consultant teams, which expedited the processing of CEs. Finally, the design team implemented an IRR analysis to assess whether matched pairs of medical consultants could interpret data reliably for the questions in the instrument.

documents relevant to a claimant's disability filing (such as the MER). The MEDIB includes an internal four-digit code that identifies the type of document inserted, including CEs. Although the MEDIB database and eView are related to one another in the information stream, there are differences between them that reflect their respective designs and functions. As described in more detail below, one challenge that the design team had to overcome was to identify the initial characteristics of each CE so that it could be assigned ("triaged") to the appropriate medical consultant.

1. Identifying CE Characteristics (Triage using Examiner Instrument)

The sample from MEDIB included a random sample of CEs from the initial and hearings levels, which created two challenges. First, because documents in eView can be opened only one at a time, the medical consultants could not identify the type of CE in the folder until the CE was opened. Second, the MEDIB contains information that is not related to a CE or to an incomplete CE, so that the examiner and medical consultants had to develop and follow rules to ensure that the CE SSA selected for the sample was included in the study.

To address these challenges, the design team developed a triage process that was implemented by the examiners to characterize the CE and identify the basic characteristics of the CE (for example, whether the CE was related to a physical or mental allegation).[35] The COMS examiner opened the folder to identify the type of CE and other basic characteristics, such as dates of CE requests.[36] The COMS examiners also identified a "file count." The information extracted by the examiner, including the file count, was then automatically transferred to the COMS and SSA management teams.[37] The COMS and SSA management teams, which included a person who understood the clinical backgrounds of the medical consultants, used the information to assign the CE to a medical consultant. The management teams generally assigned CEs to mirror the review process at the DDS level. Namely, the management teams assigned the CEs based on the broad type of CE, rather than attempting to assign the CE on a very narrow specialty.

2. Extracting Information from CEs from the Permanent Records using the eView System

All medical consultants and examiners opened cases in a double screen system to review the permanent records of claimants in the eView system. The SSA medical consultants and examiner used their regular workstations to review the CEs. The COMS medical consultants and examiners reviewed CEs in a secure location at the SSA Headquarters building in Baltimore, Maryland in accordance with SSA data security requirements for the study.

---

[35] Many cases have only one CE, but some have multiple CEs. Thus, the team had to develop rules for selecting cases so that the examiner could identify the selected CE in cases with multiple CEs. The rules for selecting cases were documented in the codebook provided to the clinical team.

[36] As a starting point for the study, SSA sent the design team 700 CEs for review, and our design team selected 600 CEs for inclusion in the IRR study. The extra 100 CEs were necessary because, as noted above, some CEs in the MEDIB included incomplete information on CEs. After the completion of the 600 CEs in the IRR study, the plan was to proceed to review 5,400 CEs (including the extra CEs that were selected in the initial review). The COMS examiner reviewed all 700 CEs, and the COMS management team identified 600 CEs from this sample that were complete. Some of the excluded 100 cases had completed CEs, and some had incomplete CE information or information not relevant to CEs..

[37] For details on how the file count was used, see Appendix B.

The medical consultants identified CEs based on the file count (described above) from the examiner. Specifically, the SSA and COMS management teams provided the medical consultants with a hard copy list of case folder identification numbers and file counts. If a medical consultant identified a file in eView that did not have the information included in their hard copy list, they were instructed to contact a support number for further assistance. Once the medical consultants verified they were reviewing the correct file, they opened the CE in eView and simultaneously extracted information on the CE using the web-based template (which appeared on a second screen).

The medical consultants and examiners extracted information from the permanent case record in eView, which should contain the majority of information on CEs. The one exception is that the permanent case record does not include information from a development section of the case folder or other information not transmitted by the state DDS.[38] We cannot precisely assess the extent to which missing information might be reflected elsewhere in the permanent case record, but we designed the template in conjunction with SSA to include items, especially medical evidence, that could be derived from the permanent case record.

C. Revisions to the Sampling Goals and Materials to Inform Data Extraction

The original sampling goals and the materials were modified substantially due to problems identified with the initial medical consultant instrument. The design team took several steps to address these issues, including conducting focus groups with medical consultants to assess the problems, developing a codebook to guide data entry, and conducted a training session. The result of this process included a more thorough documentation to guide the data extraction process, though it also led to substantial delays in the contract.

1. Reliability Problems with Initial Medical Consultant Instrument

The COMS and SSA medical consultants and examiners conducted a preliminary data extraction in the fall of 2009 using a limited sample of CEs that was ultimately stopped due to concerns in the reliability of data entry. Specifically, the COMS and SSA medical consultants extracted data for 129 CEs.[39] In reviewing these cases, the design team found that several data elements had rates of agreement below 70 percent, and most items were below 80 percent.

To identify the potential sources creating the reliability problems in the medical consultant instrument, the design team conducted focus groups with the COMS and SSA medical consultants. During the focus groups, the design team identified multiple issues with the original instrument, including problems with vague or subjective wording and items that were difficult to find in the permanent record. Additionally, several medical consultants noted fatigue in completing a very long instrument, which often took over one hour to complete.

[38] According to the COMS examiner who works with development section in his work at a DDS, the development section includes an authorization communication from the DDS to the CE provider to do the exam and the appointment letter sent to the claimant. Additionally, some DDS agencies might not transmit all of the information from their cases into the permanent record because CEs are tracked in different ways across states.

[39] At that time, there were also 81 SSA cases and 72 COMS cases that were not matched. We summarized the findings for the initial data extraction in a memo dated November 18, 2010.

2. Pretesting the Revised Medical Consultant Instrument

Following the focus groups, the design team, in conjunction with SSA, made multiple revisions to the instrument. The design team clarified the wording of several questions and eliminated other questions that were not central to SSA's research questions. The result was a more efficient instrument that the medical consultants could use to extract information regarding CEs in under an hour. The final instruments are summarized below in Chapter III.D.

In the process of revising the instruments, the COMS medical consultants conducted four pretests where pairs of COMS medical consultants extracted data for a limited set of CEs. After each review, the design team made changes to the instrument based on the findings from the feedback, which included data from the pretests and qualitative input from the COMS medical consultants on any issues interpreting the question. After the fourth pretest, the design team was satisfied that the medical consultant instrument substantially reflected adequate IRR statistics and could meet reliability thresholds (described in Chapter IV) for a larger sample.

3. Development of Codebook and Training

To facilitate reliable data extraction, the design team developed a codebook for the final medical consultant instruments, which were used the pretests described above. The purpose of the codebook was to ensure that medical consultants had a common understanding of the goals of data extraction and to address any outstanding questions they had about the review process. The codebook provided guidelines on where to find information in the electronic folder and CE, and how to code data elements. The design team revised the codebook following each pretest, and the final codebook was used during the full review by the COMS and SSA medical consultants.

The design team also led a training session for the COMS and SSA medical consultants to describe the codebook and introduce the website. Specifically, the COMS and SSA medical consultants attended a two hour in-person training led by the design team, in which they reviewed the data extraction protocols and addressed questions from the medical consultants. Following the training, the medical consultants extracted data for a single test case to ensure they had a working knowledge of the instrument and website.

The design team did not provide formal guidance or training to examiners because the examiner instrument included a limited set of objective questions. Unlike the medical consultants, the examiners opened CEs and entered responses without any formal training or guidebook. We conducted a small-scale IRR analysis of the examiner instrument to ensure that it was operating as envisioned, particularly for the triage.

4. Revised Sampling Targets Due to Delays in Data Extract

Data extraction was restarted in Fall 2011 using the final CE Review instruments (described below), but it was cut short due to an SSA administrative decision to end data collection in March 2012. The implication was that the review teams extracted information on a more limited sample of CEs than was envisioned in the design noted above.

The final sample for the medical consultant instrument included 289 adult CEs extracted by the COMS and SSA medical consultant review teams. The original target sample size included 300 CEs, which were evenly stratified by adjudication level. However, 11 child CEs were dropped following

the SSA administrative decision noted above because the child CE sample was too small for a rigorous IRR analysis.

The final sample for the examiner instrument included 25 adult CEs extracted by the COMS and SSA examiner teams. The original target sample size included 26 CEs, which were evenly stratified by adjudication level (13 CEs) each. However, 1 child CE was dropped for the same reason noted above. Unlike the medical consultant instrument, which required a full scale test of all questions, the primary goal of the examiner instrument was to assess the reliability of the triage data entry. As will be shown in Chapters V and VI, because the examiner IRR analysis showed strong agreement between SSA and COMS examiners on key items, including the type of CE, we concluded that the triage process described above was working as envisioned and did not require a more expanded sample for further testing.

5. Characteristics of Sample for Medical Consultant Instruments

Exhibit III.1 shows the CE characteristics and adjudication level of the CEs selected for the analysis of the medical consultant instrument. The CEs were roughly split between physical and mental health CEs (147 and 142 CEs, respectively) and across the initial- and hearings-level CEs (140 and 149 CEs, respectively).[40] Of the cases reviewed, 168 were for an SSDI claim, 116 were for an SSI claim, and 5 were for "other."[41] We have limited data on other administrative characteristics of the sample that were provided at the outset of data extraction, such as age and primary impairment code. For all CEs, the average age of the claimant was 45 years old and we do not find any notable differences by age. For impairment, we find some differences in the distribution of impairments. For example, initial level CEs include more claimants with mental/cognitive impairments relative to those at the hearings level. The differences in characteristics by adjudication level might reflect actual differences in the composition of CEs at the initial and hearings level (given that CEs were roughly sampled in accordance with their distribution in the overall number of CEs) or may result from differences that arose due to the sampling process for the IRR analysis. Unfortunately, we cannot assess how the selected CEs differ from those in the general population of CEs because national statistics on characteristics, such as CE type, do not exist.

[40] The slightly higher number of cases at the hearings level reflects that most of the child cases initially reviewed had been at the initial level. The sampling strategy was updated after 150 cases had been initially reviewed, almost all of which were sampled from initial-level cases.

[41] Includes other benefits (e.g., Childhood Disability Benefits and Disabled Widow or Widowers), and one case with no information on case type.

Exhibit III.1. Characteristics of the CEs Selected for the Study

	Total	Initial	Hearings
Type of CE Exam			
Mental	142	65	77
Physical	147	75	72
Internal Medicine	104	65	39
Specialty/Musculoskeletal	43	10	43
Adjudication Level			
Initial	139	139	0
Hearings	149	0	149
Claim Type[a]			
SSDI	168	94	74
SSI	116	41	75
Other	5	5	0
Age			
Mean age	44.6	46.5	42.7
Impairment[b]			
Mental/Cognitive	73	41	32
Musculoskeletal	127	58	69
Circulatory	22	15	7
Nervous	17	8	9
Endocrine	14	7	7
All Others[b]	35	10	25
Total	**289**	**140**	**149**

[a]The claim type includes SSDI (all SSDI claims and concurrent SSDI and SSI claims), SSI only (claims for SSI that do not include concurrent SSDI claims), and others. "Other" includes childhood disability benefits, disabled widowers benefit, and one CE that had missing information.

[b]All other impairments include impairments with small sample sizes in our sample, including digestive, neoplasm, respiratory, and impairments classified as "other" in SSA administrative data.

D. Final Instruments

Below, we summarize the final data extraction instruments. A detailed summary of the original SSA research questions, the separate data extraction instruments for the examiner and medical consultant reviews, and a crosswalk of the research questions with the instrument are provided in Appendix A. The codebook developed by the design team for the medical consultants is included in Appendix B.

1. Examiner Instrument

The primary objectives of the examiner review were (1) to ensure that the CE could be assigned by the management teams to the appropriate medical consultant and (2) to be brief. The design team estimated that it took approximately 15 minutes per case to complete all sections. The design team developed these instruments to extract information from the worksheet and related file documents that reflect CE processes, as well as reports that include background information on provider qualifications (e.g., provider letterhead that noted licenses). The design team also obtained additional information to address SSA study questions about processes, such as whether medical records

arrived after the CE was ordered, and qualifications of the CE provider, including his or her professional status (license and board certification status). While not specified in the regulations or Green Book, this additional descriptive information provides important insights regarding CE processes and provider qualifications that might influence the quality and completeness of the exam. The examiner instrument included the following sections, most of which were directly related to the process of ordering a CE:

- **Type of CE.** This section included questions on the type of CE exam ordered, which was divided into four general categories: (1) adult physical, (2) adult mental, (3) child physical, and (4) child mental. Adult physical CEs were further separated by medical specialty, including general medicine, cardiology, and neurology. Adult mental CEs could be classified as mental status, psychological testing, or both. The first three diagnoses and/or impairments listed by the CE provider were also noted in this section.
- **Process of Ordering a CE.** This section included information on the process and costs of ordering a CE, including information on the number of medical sources.
- **Case Dates: DDS for Initial-Level Decisions; ALJ for Hearing-Level Decisions.** This section included questions on the CE request and receipt dates to document the timeline in the CE procurement process.[42]
- **Qualifications of the CE Provider.** This section included questions such as whether the CE providers were treating sources and whether their licensure status was noted in the CE. The medical consultants entered information on the name of the medical consultant who conducted the CE. For physical health exams, while not required in the regulations, we also indicate whether the exam was performed by a physician and then identify the board certification status of the physician. We identified the certification of providers based on information from the American Board of Medical Specialties (ABMS).[43]
- **MSS from CE Provider.** This section included questions about the procedural aspects of reviewing the MSS from the CE provider, including whether the DDS worksheet or ALJ opinion requested the MSS and whether the CE authorization or invoice included this request.
- **Follow-up Contact with CE Provider.** This section included questions to determine whether the DDS needed to follow up with the CE provider for more information, such as to request additional details, to clarify or correct a finding or statement, or to obtain a signature.

[42] The website includes one additional question on the date the CE was scheduled that was added after the COMS examiner had completed the triage of cases. Hence, this information is available on the website, but not in the report.

[43] We reported information from the CE report (provider's name) and looked up their board certification status in AMBS: http://www.boardcertifieddocs.com/abms/static/home.htm;jsessionid=FC0B5977A443A36F2FB6A8DF98296AB8 (accessed August 12, 2012).

2. Medical Consultant Instrument

The 10 general medical consultant sections covered the largest portion of the data extraction and were fundamental to addressing SSA's core research questions, especially those relating to CE process and quality. The medical consultant questions also posed the greatest risk of obtaining unreliable data interpretations, given that some of them involved subjective elements, such as assessments of overall quality of the CE. The design team estimated that the approximate time to complete these questions was between 45 and 60 minutes, with slightly longer times to review hearings-level CEs with large amounts of MER. The medical consultant instrument included the following sections:

- **Worksheet Review.** This section included questions about the purchase of the CE, including whether there was any reason given for ordering it, and, if so, whether it was to obtain more recent medical evidence.
- **Medical Evidence Documentation.** This section included questions on the adequacy of the medical evidence sent to the CE provider: whether any MER was sent at all and whether the CE examiner listed at least one MER item in the CE Report.
- **Medical History—Present Illness.** This section included several questions on the history of the present illness related to chief and non-chief complaints. A major challenge in designing and ultimately reviewing data for this section was in defining a chief complaint, which can change over time in a claimant's medical history. To address this issue, the design team, in consultation with SSA, had to develop detailed rules to identify chief complaints and non-chief complaints that generally closely aligned with the claimant's primary or secondary diagnosis at application(see question I.3 in Appendix Exhibit A.3).[44]
- **Additional History.** This section included questions about information on additional claimant medical history, such as medications taken, whether family and/or past medical history were noted, and whether the review of systems was documented.
- **Physical Exam Findings.** This section, which was completed for adult and child physical CEs, included questions about the physical exams, such as whether vital signs were recorded and whether any part of the exam was recorded on a standardized form.[45]

---

[44] According to SSA staff, when a disability claim is filed, the SSA-3368/3369 requires the claimant to list allegations of his impairment. At the initial and reconsideration levels, the SSA-831 is the form used to document the primary and secondary diagnoses used to make the disability determination. For hearings level cases, either the SSA-831 or ALJ's decision is used to document the primary and secondary diagnoses used to make the decision. However, the concept of the "chief complaint" must include a recognition that the claimant's allegations and the diagnoses reached in decision making may change over time. As medical evidence is obtained and CE exams or tests are performed, the claimant's original allegations are not what the agency's decisions may be based on. This creates a challenge in data extraction as the team must develop some rules to identify a concept of chief complaint for CEs. The design team outlined a seven step process that the medical consultants could follow to use these terms (primary and secondary diagnoses) to describe "chief complaint". The definition generally aligned with the primary and secondary diagnosis on the 831 and, in the cases of hearings level CEs, the ALJ's opinion.

[45] We processed some specific exam types not noted below for adults and children (such as cardiovascular or skin diseases for adults and children, and multiple body systems or growth or immune systems for children) as part of the generalist exam.

The ensuing sections of the physical exam were completed for specific subgroups. The largest subgroup was covered in the second section, Generalist Exams, which applied to adult and childhood physical exams in general medicine/internal medicine/family medicine, cardiology, pulmonary, rheumatology, gastroenterology, hematology/oncology, endocrinology, genitourinary, skin diseases, and other specialties. The questions in this subsection included more details on specific physical attributes, such as joint findings and abdominal characteristics. The third section referred specifically to musculoskeletal and orthopedic exams and included nine questions. The fourth section included questions on neurology CEs (as well as speech language pathology and neurosurgery CEs). The fifth section included questions on ophthalmology exams. The sixth and final section included questions on ear, nose, and throat attributes.

- **Mental Health.** This section applies to child and adult mental health exams and includes questions such as whether the CE assessed the claimant's thought processes and content, mood, cognition, and judgment.
- **Lab Studies/X-Rays/Tests.** This section includes questions about lab studies, X-rays, and other tests either performed by or available to the CE provider, as well as what types of tests were performed and whether they were in compliance with the Listing of Impairments criteria.
- **CE Report Assessment by Medical Consultant.**[46] Questions in this section ask the medical consultants to assess the overall CE based on the information collected in the previous sections. It includes their assessment of key medical information, such as whether the CE report included a reasonably stated diagnosis and prognosis, and their opinions of the CE findings.
- **MSS Involving Functional Capabilities or Childhood Domains (Adults/Children)**. This section includes questions on adult and child functionality, beginning with whether the CE provided an MSS, and then separate sections for adult and child exams about the MSS related to functions. For adults, the questions relate to the ability to stand, walk, travel, and so on (as well as social functioning and adaptation for adult mental health CEs). For child CEs, this section addresses functional impairments along several domains (e.g., completing tasks).
- **Medical Consultant Assessment of Overall Completeness of CE Report.** This section includes questions about the overall completeness of the exams. The questions include information on whether CE can be used to support an disability determination and whether SSA received all the information that was ordered.

[46] In the medical consultant instrument, this section is referred to as CE Report by Medical Consultant, though we changed the name here to clarify that the opinions are provided by the review team (as opposed to a DDS medical consultant).

IV. METHODOLOGICAL APPROACH TO ASSESSING RELIABILITY

In this chapter, we summarize our approach for studying the consistency of responses across SSA and COMS medical consultants and examiners. We begin by outlining the general goals of an IRR analysis and discuss its importance for this study. We then describe our thresholds for reporting statistics and conclude with an overview of how we present the IRR analysis findings in the ensuing chapters.

A. Purpose of IRR

We assessed the reliability for the following samples:

- **Medical consultant review (289 CEs)**
- **Examiner review (25 CEs**)

The purpose of the IRR analysis was to evaluate the utility of the CE template as a data extraction tool. Given the subjectivity inherent in evaluating CEs, the IRR analysis provides a critical measure of whether consensus between medical consultants and examiners has been reached and whether the template can provide reliable data on CE study questions for the contract. The IRR findings from the early part of the study, which revealed reliability problems, underscore the importance of testing the reliability. Although the IRR analysis is essential in assessing the likelihood that multiple medical consultants and examiners score items in the instrument in comparable ways, it does not provide a measure of the validity of the template (that is, how close assessments are to a "true" value). Because COMS and SSA medical consultants and examiners generally had comparable experiences in reviewing CEs, we do not know which medical consultant or examiner is more accurate when there is disagreement about how to code a particular item. Nevertheless, this analysis provides an important indication of the quality of the instrument's items and its utility as a data extraction tool. We used the early IRR information to revise the instrument and refine protocols for data extraction. In the future, SSA can use the IRR analysis presented in this report to further refine data extraction and/or as a baseline for testing how future medical consultants and examiners compare to those who participated in this study.

B. IRR Statistics Thresholds

We use a percentage agreement measure to track the reliability of data entry. Below, we describe how we develop the percentage agreement measures and highlight the thresholds we established to assess the reliability of each question.

1. Measuring Percentage Agreement

We use the percentage agreement as our primary metric for assessing the reliability of the instrument questions. The percentage agreement represents the exact agreement for each question between rating pairs. For example, a 70 percent agreement would show that COMS and SSA medical consultants and examiners exactly agreed on the responses in 70 percent of the cases. For the medical consultant review, we show the percentage agreement for the 289 CEs included in the paired review for this group. For the examiner review, we show the percentage agreement for the 25 CEs included in the paired review for this group.

Where possible, we attempt to group the response options into dichotomous categories to present percent agreement (e.g., yes/no). Most of the questions in the examiner and medical consultant sections of the template already had dichotomous response options, so this required no change in most cases. For items with multiple response options, we only collapsed variables that had reasonable analytic groups (for example, when the question contained two "yes" categories, we collapsed them into one "yes" category). In some instances two "yes" options made the question easier for reviews to code.[47]

Finally, for some questions, particularly in the examiner section of the template, several variables had response options that were continuous or multi-level responses (such as number of medical sources, fees, and schedule dates). In these cases, we show both percentage agreement, which shows exact matches, and correlation coefficients, which show strength and direction of the linear relationship of the continuous variables. For example, the examiners who collect fee information might report fees of $150 and, say, $151. In this case, there would be no agreement, but it is likely that the small difference between the ratings could still be highly correlated, which has implications for producing IRR statistics for continuous variables (described below).

A major advantage of showing percentage agreement is that it allows for comparability across the majority of items collected by the medical consultants and examiners for presentation purposes. Specifically, this metric provides a straightforward method showing which questions and sections are consistently most reliable. In the text, we show the summary dichotomous variables, but for completeness, in Appendix C we show the full summary of all responses for all questions in the larger medical consultant IRR analysis.[48]

2. Reliability Thresholds

Although there is no universally accepted threshold for acceptable agreement, we apply a guideline of 70 percent agreement or greater for questions with dichotomous response categories to evaluate the quality of IRR. For the limited number of continuous variables in our analysis, we present percent agreement and correlation coefficients. We establish a threshold of 0.90 correlation coefficient or higher for continuous variables.

This reliability threshold is consistent with rates noted in the previous literature for this type of data extraction. As documented in Chapter III.C, the task of locating and coding data proved more challenging than initially thought. Specifically, we discovered that the diversity of documents reviewed by the COMS/SSA pairs posed a major challenge, given the lack of consistent,

[47] For example, questions about whether the CE provider referred to Medical Records as a group, offered two options for "yes," ("yes referred to and provided MER" or "Yes, referred to, but did not provide MER). However, in analyzing this question, the distinction we are primarily concerned with is simply whether or not the CE provider in this case reviewed the item as a group regardless of individual MER items.

[48] We also report kappa statistics in Appendix C, which are useful in assessing the reliability of questions with multiple response options. The kappa statistic provides a useful estimate of the degree of consensus between two judges after correcting for the amount of agreement that could be expected by chance alone. However, for items with a high prevalence of "yes" responses, kappa is difficult to interpret. Because many items in the template have this characteristic, we focus on percentage agreement and note that the kappa statistics should be interpreted with caution. For this reason, we restrict the presentation of kappa statistics to Appendix C. One set of widely used thresholds for kappa is as follows: < 0 is no agreement, 0–0.20 is slight agreement, 0.21–0.40 is fair, 0.41–0.60 is moderate, 0.61–0.80 is substantial, and 0.81–1.0 is almost perfect agreement (Landis and Koch 1977).

standardized review forms across states. For example, for the pairs to agree, they needed to have reviewed the same documents and the same parts of the documents to locate the key information required to make a judgment. In addition, although some of the items require straightforward judgments, others require the medical consultants to locate and analyze information across parts of different documents within the same CE and draw inferences to make their judgments. Several previous studies have used similar types of thresholds in extracting data that had challenges in identifying or assessing key items (see especially Stemler (2004) who conducted a broad review of studies that relate most closely to this study; see also Perez-Johnson et al. 2009 and Raudenbush and Sadoff 2008 for related studies).

3. Presentation of Findings: Fair, Moderate and High Agreement

We report IRR statistics for each question that had a paired review by COMS and SSA. For the medical consultant samples, there were 287 pairs. For examiners, there were 25 pairs. Several questions have skip patterns in which the question was asked contingent on the response to a previous question. Other items were asked only for a subsample of CEs. In both these instances, the sample for the rating pairs was small (that is, less than 287 in the medical consultant sample and less than 25 in the examiner sample). For reference, the skip patterns for each question are documented in Appendix Exhibits A.2 and A.3.

We present findings for each question for which a matched pair of COMS and SSA medical consultants and examiners provided a response. For each question, we present the distribution of responses and percentage agreement. We use the percentage agreement to assess whether the question meets our 70 percent reliability threshold. The distribution of responses provides a context for how many CEs contain information on each variable. This context is especially useful in considering questions about information that might be easy to identify, which drive up the percentage agreement. For example, a question that is predominantly answered "yes" (or "no") might be more straightforward to identify, such as whether medical information was in a narrative format.

It is important to note that some questions might have the same or similar distributional responses but different percentage agreement. This issue arises because the percentage agreement measures perfect agreement across responses, whereas the distributional responses show the aggregate responses of the medical consultants and examiners. To illustrate at the extreme, suppose there are four questions with categories of "yes" and "no." It is possible for the SSA medical consultant to answer "yes" to two questions that the COMS medical consultant answers with "no," and vice versa. In this case, there is no agreement, but the distributions of CEs match.

For questions that meet our 70 percent threshold, we provide an additional qualitative rating of fair (70 to 79 percent), moderate (80 to 89 percent), or high (above 90 percent) to depict broad trends for the 142 questions discussed in the next chapter. This classification provides insights into relative strength of the reliability. Similar types of qualitative assessments have been used in previous IRR analyses (e.g., see Hartling et al. 2012). However, because of the arbitrary nature of these labels, some caution should be used in interpreting these categories as strict cutoffs. For example, there is very little difference between questions rated at 90 percent agreement and those rated at, say, 89 percent, though the former would be categorized as high and the latter as moderate in this framework. Nonetheless, these labels are helpful in categorizing the reliability trends for the large number of questions included in the examiner and medical consultant instruments.

Finally, we present a summary statistic in each exhibit for the percentage agreement of all questions within a section and identify questions that do not meet the thresholds described above. The summary statistic is a general estimate of reliability within a section, which provides useful context in describing the questions and trends. For example, we expect that questions in sections with more subjective information, such as medical consultant assessments of CE quality, would have lower overall percentage agreement than questions in sections that contain predominantly objective information, such as type of CE exam. We present a summary of questions that do not meet our thresholds for further follow-up. This summary provides useful context for questions that were less reliable, and hence have limitations for characterizing CEs. In the future, SSA might choose to revise these questions, eliminate them, or simply accept that they are measured with error.

V. SUMMARY FINDINGS FROM IRR ANALYSIS

In this chapter, we summarize our overall findings for the instrument. We present summaries of the percentage agreement by section to illustrate the general patterns of reliability. This summary provides context for the detailed question by question summary that appears in Chapter VI. We also identify the 14 of the 142 questions from the examiner and medical consultant instruments that fell below our 70 percent threshold. SSA could target these 14 questions for future revision should they use the web-based template in future development efforts. Finally, we conclude with a sensitivity test of our findings that shows the patterns of agreement are similar across the initial and hearings level, indicating that the reliability of data extraction did not differ substantially for a key subgroups of analytic interest.

A. Overall Agreement for Instruments Was 85 Percent

The overall percentage agreement across all sections in the final medical consultant and examiner instruments was 85 percent, well above the 70 percent threshold established for the study (**Exhibit V.1**). Of the 142 questions in the instrument, 128 questions exceeded the reliability thresholds, including 99 that were above moderate agreement (i.e., above 80 percent agreement). Across sections, the percentage agreement ranged from 78 percent (Medical History Section) to 95 percent (Type of CE, Case Dates, and Qualifications of the Providers).

The percentage agreement was generally higher in the medical consultant sections that contained relatively objective information on the CE process. The combination of questions in the physical exams (general and specialty-specific), mental health exams, and lab studies/exams/tests sections all had over 87 percent overall agreement ratings.

The overall percentage agreement was generally lower in the medical consultant sections where the medical concepts were not well defined in federal regulations. For example, the medical history and medical evidence documentation sections both had overall agreement slightly below 80 percent. However, even in these sections, the majority of questions exceeded our reliability threshold of 70 percent.

Some caution should be used in assessing the reliability of the examiner sections given the limited sample of 25 CEs. In general, questions on type of CE, case dates and qualifications had high agreement. As will be shown in Chapter VI, the questions on the triage process had near perfect agreement, which provided a strong assessment that the triage process was working as intended. However, the questions on other processes, such as the number of medical sources, did not meet the reliability thresholds.

B. 14 Items Fell Below Thresholds

Of the 14 items that fell below the 70 percent threshold, 13 were for a subgroup of CEs, many of which had very limited samples (**Exhibit V.2**). In six cases (B.3, D.1, D.2, D.3, D.4, and K.3c1), the sample size was 13 or fewer CEs, so any concerns over reliability for these questions should be considered exploratory given the limited samples. Of the remaining eight CEs, the less reliable questions tended to fall in one of the following groups: documenting the process of obtaining an MSS, which might not have been well documented at the DDS in an invoice or worksheet (C.3, H.1 and H.2); obtaining medical information on a difficult-to-document complaint (I.3d, I.4c, and I.4d),

or subjective medical assessments by the medical consultant that were based on limited information (J.7b, N.4, and N.11).

If SSA considers future data extraction efforts, the agency can choose to modify these questions to improve reliability or simply not change them. If SSA does not change the questions, it should view the responses as being measured with error.

C. Limited Differences in Reliability of Questions by Adjudication Level

To assess whether differences exist in the reliability of data entry by adjudication level, we generated tabulations for all of the questions in the medical consultant instrument by adjudication level. One hypothesis was that because hearings level CEs tend to include more medical content relative to the initial level CEs, it could make it more difficult for the medical consultants to reliably extract information on hearings level CEs. Hence, our comparison of this subgroup provides an important sensitivity test.[49]

Overall, we find very small differences in the percentage agreement by adjudication levels, as the overall agreement for all questions at the initial level was 86 percent and 85 percent at the hearings level (data not shown).[50] These findings indicate that the medical examiner instrument was consistently reliable across the initial and hearings levels.

[49] Throughout the pretest, we also assessed whether differences exist by mental health and physical health CEs. We did not find any notable differences.

[50] There are only two questions where the percentage agreement for both levels either do not both exceed 70 percent or fall below 70 percent (C2 and I4c). Specifically, the percentage agreement for C.2 (rationale for the worksheet) was 67 percent at the initial level and 81 percent at the hearings level. We continue to assess this variable as exceeding the 70 percentage agreement for the overall (i.e., hearings and initial level) and will include it in our second report for the study. The percentage agreement for I4c was 74 percent at the initial level and 60 percent at the hearings level. We will not include this variable in our second study given that, as noted above, it did not excel the 70 percent agreement for the overall sample.

Exhibit V.1. Summary Agreement Across Medical Consultant and Examiner Sections

	Percentage Agreement for Rating Pairs in Section	Number of Questions	Did Not Exceed Reliability Threshold[a]	Questions That Exceed Reliability Thresholds[a]			
				Total	Fair	Moderate	High
Medical Consultant Sections							
Worksheet Review	82.5	3	1	2	1	0	1
Medical Evidence Documentation	79.8	2	0	2	1	1	0
Medical History and Present Illness	78.1	14	3	11	6	4	1
Additional Medical History	88.9	9	1	8	0	3	5
Physical Exam Findings	87.1	7	0	7	2	2	3
Physical Exam Findings: Generalist Exams[b]	88.6	22	0	22	6	3	13
Physical Exams: Orthopedic/Musculoskeletal Exam	87.8	9	1	8	2	1	5
Mental Health Exams	93.2	11	0	11	0	2	9
Lab Studies/Exams/Tests	89.7	5	0	5	1	2	2
CE Report Assessment by Medical Consultant	81.5	22	2	20	5	8	7
Medical Source Statement and Functional Capacities	82.7	15	0	15	3	10	2
Overall Completeness of CE Report	85.3	3	0	3	0	2	1
Examiner Sections							
Type of CE, Case Dates, and Qualifications of the Providers	98.1	11	1	10	0	0	10
Process of Obtaining a CE, Medical Source Statement from CE Provider, and Follow-Up with Provider	64.3	9	5	4	2[c]	0	2
Total							
Overall	84.7	142	14	128	29	38	61

Source: COMS and SSA Medical Consultant and Examiner CE Review data.

Notes: Authors' calculations based on percentage agreement from Exhibits V.1-V.14. The percentage agreement shows the exact agreement between COMS and SSA examiners. The second column shows the number of questions in each section that met the 70 percent agreement threshold relative to the number of questions in that section. The qualitative ratings represent percentage agreement that was "below" threshold (below 70 percent or a correlation coefficient below. 90 for continuous variables), "fair" (70 to 79 percent), "moderate" (80 to 89 percent), or "high" (above 90 percent) to depict broad trends for each section.

[a] The thresholds for percentage agreement were 70 percent with dichotomous variables and 90 percent for continuous variables.

[b] There were 17 questions in this section, but one question had six categories (check all that apply), so we report responses separately for each category.

[c] We report two questions that had continuous responses as "fair" that met the correlation threshold (of at least 90 percent), but did not meet the percentage agreement threshold.

Exhibit V.2. Summary of Questions That Fall Below Reliability Thresholds

Question		Response Options	Rating Pairs	SSA	COMS	Percentage Agreement
B.3	For adult mental only, what type of exam was received?	Mental Status examination/ Psychological testing/Both	13	10/2/1	8/0/5	69.2%
C.3	Did the Worksheet note that the CE was ordered to obtain more recent evidence?	Yes/No	27	15/12	13/14	55.6%
D.1	How many Medical Sources (MSs) were identified on the 3368 or 3820?	Number	12	3.9	4.3	66.6%
D.2	How many MSs provided medical information? (There may be more in the file than are listed on the 3368/3820.)	Number	12	3.5	6.7	33.3%
D.3	Number of MSs providing medical information before initiating the CE?	Number	12	3.2	6.3	33.3%
D.4	How long did the disability examiner wait (after the last request for MER) before purchasing the CE?	Less than 21 days or 1 month/More than 1 month	12	2/10	7/5	58.3%
H.1	Did the DDS Worksheet or ALJ's opinion note that an MSS was expected/requested?	Yes/No	25	2/23	9/16	56.0%
H.2	Did the CE authorization or invoice request an MSS?	Yes/No	25	8/17	16/9	52.0%
I.3d	Was anything that made the chief complaint-related medical condition better (including treatment) or worse described?	Yes/No	238	132/106	161/77	65.1%
I.4c	Was the approximate time of onset of at least one non-chief complaint allegation or possible impairment described?	Yes/No	164	97/67	81/83	67.1%
I.4d	Was at least one factor that contributed to ongoing worsening or improvement of at least one non-chief complaint allegation or impairment?	Yes/No	164	83/81	78/86	67.7%
J.7b	Was the family medical history (FMH) pertinent to the claimant's allegations noted?	Yes/No	142	84/58	61/81	68.3%
K.3c1	If abnormal, was SLR/tension signs confirmed in another body position?	Yes/No	8	4/4	5/3	62.5%
N.4	Were all allegations that were evaluated or listed by the provider previously known to SSA?	Yes/No	289	224/65	205/84	66.4%
N.11	Do you agree with the ALJ that the MER was not sufficient to support a case decision without your current CE?	Yes/No	149	105/44	134/15	65.8%
Summary for Section						
All Responses						65.2%

Source: COMS and SSA Medical Consultant and Examiner CE Review data.

Notes: The findings include all questions from Exhibits V.1-V.14 that fell below the 70 percent agreement threshold. The rating pairs column indicates how many medical consultants from both teams reviewed the question. The summary for "All Responses" includes the percentage agreement for all rating pair responses for every question in the section. The qualitative ratings represent percentage agreement that was "below" threshold (below 70 percent), "fair" (70 to 79 percent), "moderate" (80 to 89 percent), or "high" (above 90 percent) to depict broad trends for each question.

VI. DETAILED FINDINGS FROM IRR ANALYSIS

In this chapter, we present a detailed summary of the findings the 289 medical consultant CE paired reviews and the 25 examiner CE paired reviews. We first present findings from the medical consultant IRR analysis, which comprises the majority of the questions in the CE Review. For each of the 10 sections within the medical consultant instrument for the adult CEs covered in our sample, we present an exhibit that summarizes the reliability of individual questions and the overall reliability of the section. To provide a further qualitative assessment of reliability, we categorize the percentage agreement for the individual questions who meet our reliability thresholds into the three categories outlined in Chapter IV: "fair" (70 to 79 percent), "moderate" (80 to 89 percent), and "high" (above 90 percent). We then present a brief summary of findings from the examiner instrument. Given the limited sample sizes for the examiner analysis, some caution should be used in interpreting the results.

A. Medical Consultant Sections

Below, we review the findings from the 289 matched pairs of medical consultants for the 10 medical consultant sections. We present a separate exhibit for each section of the medical consultant instrument. The one exception is that for the physical exam section we present an exhibit for the overall physical exam and two exhibits that cover specialties within adult physical CEs.

1. Worksheet Review

The overall percentage agreement for the worksheet section was 83 percent (**Exhibit VI.1**). Although medical consultants were consistently able to find the DDS worksheet in the folder, they had some difficulty in identifying specific details of a particular claim. One challenge in identifying these details was that the format of worksheets varied substantially by state (see Chapter II for more details).

Of the three questions in the section, two exceeded the 70 percent threshold. There was high agreement for whether a worksheet was in the case folder (C.1), which medical consultants found in the vast majority of CEs. However, the reason for ordering a CE (C.2), which included 256 pairs of medical consultants who identified a worksheet, had fair reliability. Additionally, for the 27 paired CEs who responded "yes" to C.2, the question about whether the worksheet noted that the CE was ordered to obtain more recent evidence (C.3) did not meet the 70 percent threshold. During the pretests, the design team made several alterations to obtain more specific content from the worksheet, particularly in questions C.2 and C.3. However, the fair reliability shown for question C.2 and low reliability for C.3 illustrate the challenges of obtaining information from worksheets that vary substantially across states.

2. Medical Evidence Documentation

The overall percentage agreement for the medical evidence documentation section was 80 percent (**Exhibit VI.2**). This section included two questions that referred to the CE provider's review of forwarded medical records; once found, this evidence could be objectively reviewed to determine whether the CE provider had reviewed previous medical records.

Exhibit VI.1. Worksheet Review

Question		Response Options	Rating Pairs	SSA	COMS	Percentage Agreement	Qualitative Rating
C.1	Was a DDS worksheet in the E-file?	Yes/No	289	270/19	262/27	93.1%	High
C.2	Was any reason given on your worksheet for ordering your CE?	Yes/No	256	61/195	61/195	73.4%	Fair
C.3	Did the worksheet note that the CE was ordered to obtain more recent evidence?	Yes/No	27	15/12	13/14	55.6%	Below
Summary for Section							
All Responses						82.5%	

Source: COMS and SSA Medical Consultant CE Review data

Notes: The sample includes selected CEs rated jointly by SSA and COMS medical consultants. In total, the COMS and SSA team jointly reviewed 289 adult CEs, which were stratified by type of CE exam (mental versus physical) and determination level (initial versus hearings). The rating pairs column indicates how many medical consultants from both teams reviewed the question. The number of rating pairs can be less than 289 if only a subgroup group answered a question. The percentage agreement shows the exact agreement between COMS and SSA medical consultants. The summary for "All Responses" includes the percentage agreement for all rating pair responses for every question in the section. The qualitative ratings represent percentage agreement that was "below" threshold (below 70 percent), "fair" (70 to 79 percent), "moderate" (80 to 89 percent), or "high" (above 90 percent) to depict broad trends for each question.

Exhibit VI.2. Medical Evidence Documentation

Question	Response Options	Rating Pairs	SSA	COMS	Percentage Agreement	Qualitative Rating
G.1 Did the CE provider refer to medical records as a group or names of individual items in the CE report?	Yes/No	289	153/136	147/142	77.2%	Fair
G.2 Did the CE provider list at least one item of MER he/she reviewed in the CE Report?	Yes/No	96	77/19	73/23	87.5%	Moderate
Summary for Section						
All Responses					79.8%	

Source: COMS and SSA Medical Consultant CE Review data.

Notes: The sample includes selected CEs rated jointly by SSA and COMS medical consultants. In total, the COMS and SSA team jointly reviewed 289 adult CEs, which were stratified by type of CE exam (mental versus physical) and determination level (initial versus hearings). The rating pairs column indicates how many medical consultants from both teams reviewed the question. The number of rating pairs can be less than 289 if only a subgroup group answered a question. The percentage agreement shows the exact agreement between COMS and SSA medical consultants. The summary for "All Responses" includes the percentage agreement for all rating pair responses for every question in the section. The qualitative ratings represent percentage agreement that was "below" threshold (below 70 percent), "fair" (70 to 79 percent), "moderate" (80 to 89 percent), or "high" (above 90 percent) to depict broad trends for each question.

Both questions in the section exceeded the 70 percent threshold. The agreement for whether the CE provider referred to medical records (G.1) was fair. The fair agreement may be related to potential confusion in reading the instructions whether they had to rate "any reference to medical records" as "yes", even in cases where the CE provider had no medical evidence to review. For the 96 medical consultant pairs that identified reviewed medical records as a group (G.2), there was moderate agreement on whether the CE provider also listed at least one specific MER item.

3. Medical History and Present Illness

The overall percentage agreement for the medical history section was 78 percent (**Exhibit VI.3**). This section included a combination of questions regarding chief complaints and other complaints. In general, for the questions of whether complaints existed, the medical consultants had higher percentage agreement ratings, but the ratings fell when the medical consultants categorized specific aspects of these complaints, especially the time of onset and factors that contributed to the improvement or worsening of a complaint. The medical consultants also had stronger agreement in identifying a chief complaint (I.3 series) using the study definition, but agreement rates were relatively lower for non-chief complaint questions (I.4 series).

Of the 14 questions in this section, 11 exceeded the 70 percent threshold. The three general questions about the medical history (I.1, I.2, and I.6) exceeded the 70 percent threshold. The medical consultants determined whether the history was in narrative format (I.6) with high agreement, whether there was a comment in the medical history about its reliability with moderate agreement (I.2), and whether the provider indicated who gave the medical history (I.1) with fair agreement. Four of the five questions about the chief complaint (I.3 series) exceeded the 70 percent threshold. The medical consultants identified a chief complaint (I.3) with moderate reliability. For the 238 pairs who identified a chief complaint, there was moderate agreement on two questions related to clarifying and documenting the severity of the chief complaint (I.3a and b). For the other two chief complaint questions, reliability was lower: it was fair for documenting the time of onset (I.3c), but the factors contributing to the chief complaint (I.3d) scored below the 70 percent threshold. For the five questions related to non-chief complaints, there was relatively lower percentage agreement in comparison to the chief complaint questions, as two fell below the 70 percent threshold. For the 164 pairs who identified a non-chief complaint(s), reliability was mixed in ways that were similar to the chief complaint questions above. It was fair for the clarification (I.4a) and severity (I.4b) of a non-chief complaint issue(s). However, the final two questions for non-chief complaints related to time of onset (I.4c) and contributing factors to worsening conditions (I.4d) had reliability just below 70 percent. Finally, one question that captured whether there was a history of diagnostic and/or treatment experiences for any complaint (I.5), which applied to both chief and non-chief complaints, had fair reliability.

4. Additional Medical History

The overall percentage agreement for the additional medical history section was 89 percent (**Exhibit VI.4**). The relatively strong percentage agreement likely reflects that many of the questions included readily identifiable objective information, such as prescription drug use.

Of the 10 questions in this section, nine exceeded the 70 percent threshold. The seven questions for additional medical history related to all 289 CEs; they had either moderate or high agreement. The questions with high agreement included a listing of medications, drug dosage regimens, a possible history of illicit substance or alcohol use, and use of a standardized form (J.2, J.2a, and J.3, and J8). The questions for the adequacy of the past medical history and school/work

Exhibit VI.3. Medical History and Present Illness

Question		Response Options	Rating Pairs	SSA	COMS	Percentage Agreement	Qualitative Rating
I.1	Did the CE provider specifically indicate in a separate comment who gave the medical history?	Yes/No	289	80/209	64/225	77.2%	Fair
I.2	Was there a comment in the CE report about the reliability of the medical history?	Yes/No	289	80/209	47/242	81.0%	Moderate
I.3	Per study definition, was there a chief complaint?	Yes/No	289	259/30	250/39	88.6%	Moderate
I.3a	Was the chief complaint clarified (differential diagnosis explored or a diagnosis confirmed)?	Yes/No	238	205/33	224/14	84.5%	Moderate
I.3b	Was any information provided that reflected on the severity of the chief complaint-related medical condition?	Yes/No	238	214/24	212/26	83.2%	Moderate
I.3c	Was the approximate time of onset of the chief complaint-related medical condition described?	Yes/No	238	185/53	163/75	75.6%	Fair
I.3d	Was anything that made the chief complaint-related medical condition better (including treatment) or worse described?	Yes/No	238	132/106	161/77	65.1%	Below
I.4	Were there any allegations or complaints possibly related to any medical condition, diagnosis, impairment, or process that was not related to the chief complaint?	Yes/No	289	187/102	221/68	72.3%	Fair
I.4a	Was at least one other allegation not related to the chief complaint clarified (differential diagnosis explored or a diagnosis	Yes/No	164	128/36	149/15	76.2%	Fair
I.4b	Was any information provided that reflected on the severity of at least one non-chief complaint allegation or possible impairment?	Yes/No	164	124/40	141/23	78.7%	Fair
I.4c	Was the approximate time of onset of at least one non-chief complaint allegation or possible impairment described?	Yes/No	164	97/67	81/83	67.1%	Below
I.4d	Was at least one factor that contributed to ongoing worsening or improvement of at least one non-chief complaint allegation or impairment better (including treatment) or worse described?	Yes/No	164	83/81	78/86	67.7%	Below
I.5	Was there a history of inpatient/outpatient diagnostic/treatment experiences related to the chief complaint-related medical condition or to a non-chief complaint allegation or possible impairment?	Yes/No	289	221/68	247/42	73.0%	Fair
I.6	Was at least part of the medical history described in narrative format?	Yes/No	289	282/7	276/13	93.1%	High
Summary for Section							
All Responses						78.1%	

Source: COMS and SSA Medical Consultant CE Review data.

Notes: The sample includes selected CEs rated jointly by SSA and COMS medical consultants. In total, the COMS and SSA team jointly reviewed 289 adult CEs, which were stratified by type of CE exam (mental versus physical) and determination level (initial versus hearings). The rating pairs column indicates how many medical consultants from both teams reviewed the question. The number of rating pairs can be less than 289 if only a subgroup group answered a question. The percentage agreement shows exact agreement between COMS and SSA medical consultants. The summary for "All Responses" includes the percentage agreement for all rating pair responses for every question in the section. The qualitative ratings represent percentage agreement that was "below" threshold (below 70 percent), "fair" (70 to 79 percent), "moderate" (80 to 89 percent), or "high" (above 90 percent) to depict broad trends for each question.

Exhibit VI.4. Additional Medical History

Question		Response Options	Rating Pairs	SSA	COMS	Percentage Agreement	Qualitative Rating
J.1	Was a Review of Systems documented?	Yes/No	147	90/57	88/59	87.8%	Moderate
J.2	Were any medications listed anywhere in the CE report?	Yes/No/No medication was being taken	289	235/17/37	242/21/37	90.0%	High
J.2a	Was at least one dose regimen noted?	Yes/No	229	90/138	93/136	92.6%	High
J.3	Did the CE provider inquire about a history of use of alcohol and/or illicit substances?	Yes/No	289	248/41	243/46	94.8%	High
J.4	Was the past medical history noted?	Yes/No	289	252/37	267/22	86.5%	Moderate
J.6	Was the work/school history (in the history of present illness, past medical history, or a separate section) sufficient?	Yes/No	289	217/72	249/40	86.2%	Moderate
J.7a	Was the family medical history noted? (physical (exam)	Yes/No	147	70/77	72/75	95.9%	High
J.7b	Was the family medical history pertinent to the claimant's allegations noted? (mental exam)	Yes/No	142	84/58	61/81	68.3%	Below
J.8	Was any part of the medical history recorded on a standardized form?	Yes/No	289	16/273	20/269	91.0%	High
Summary for Section							
All Responses						88.9%	

Source: COMS and SSA Medical Consultant CE Review data.

Notes: The sample includes selected CEs rated jointly by SSA and COMS medical consultants. In total, the COMS and SSA team jointly reviewed 289 adult CEs, which were stratified by type of CE exam (mental versus physical) and determination level (initial versus hearings). The rating pairs column indicates how many medical consultants from both teams reviewed the question. The number of rating pairs can be less than 289 if only a subgroup group answered a question. The percentage agreement shows the exact agreement between COMS and SSA medical consultants. The summary for "All Responses" includes the percentage agreement for all rating pair responses for every question in the section. The qualitative ratings represent percentage agreement that was "below" threshold (below 70 percent), "fair" (70 to 79 percent), "moderate" (80 to 89 percent), or "high" (above 90 percent) to depict broad trends for each question.

history (J.4 and J.6) had moderate agreement. The remaining three questions in this section were specifically for the subgroup of 147 adult physical CEs, 2 of which exceeded the 70 percent threshold. The Review of Systems (J.1) and family history for the 147 physical exams (J.7a) questions had moderate to high reliability agreements, respectively. The question about whether the family medical history was pertinent to the allegations (J.7b), which included 142 mental CEs, had reliability below our 70 percent threshold. Unlike the other questions in this section, J.7b required a more subjective assessment of the use of the family medical history, which likely explained its lower percentage agreement.

5. Physical Exam Findings

The physical exam section included a general set of questions that applied to all physical CEs (such as vital signs) and subgroups of specialty exams. An important issue in summarizing findings for these exams was that the medical consultants had to consider whether the finding was germane to the allegations being evaluated. For example, for a cardiac allegation, "sensation" did not need to be described with the same amount of detail as when the claimant alleged peripheral neuropathy. When the particular finding was relevant to the case, enough detail needed to be provided so that an examiner could determine if an impairment was severe. Below, we report the findings for the physical exams, including those for all physical CEs, the generalist physical exams, and the orthopedic/musculoskeletal exams. The sample for the adult physical CEs was 147. These CEs were split across three specialty groups: generalist (104 CEs), orthopedic/musculoskeletal (38 CEs), and neurology (5 CEs). Because of the small sample of neurology CEs, we do not include a separate exhibit with the details of the physical exam for neurology CEs (though these CEs are included in the general exam).

a. General Physical Exam

The overall percentage agreement for the general physical exam was 87 percent (**Exhibit VI.5**). The relatively strong agreement for this section reflects that most items from the physical exam could be objectively rated, such as vital signs, height, and weight.

All seven of the questions for all physical general exams exceeded the 70 percent threshold. Three items had high percentage agreement, including verifying the claimant's identification, recording of vital signs, and recording of weight and height (K.1a, K.1b, and K.1f). Two physical exam items for gait/station and the use of assistive device (K.1c and K.1d) had moderate reliability. The one apparently more subjective item related to the claimant's ability to dress (or show fine motor skill [K.1e]), which was the most frequently reported "no" item in the physical exam, had fair reliability. Finally, there was fair agreement on whether any part of the physical exam was recorded on a standardized form (K.1o).[51]

b. Physical Exam: Generalists

The overall percentage agreement for the generalist physical exam section was 89 percent (**Exhibit VI.6**). The percentage agreement in this section was impressive given the long battery of questions. The design team developed questions for the generalist physical exam section that

[51] While the standardized form question passed reliability thresholds, the percentage agreement was lower than the level for whether standardized forms were used in other parts of the CE.

Exhibit VI.5. Physical Exam Findings

Question		Response Options	Rating Pairs	SSA	COMS	Percentage Agreement	Qualitative Rating
K.1a	Was there a comment that the claimant's identification was verified at the CE?	Yes/No	147	13/134	16/131	93.9%	High
K.1b	Was pulse rate, blood pressure, and/or respiratory rate recorded?	At least 1 item/No	147	132/15	131/16	99.3%	High
K.1c	Was station or gait described?	Yes/No	147	132/15	140/7	89.1%	Moderate
K.1d	Was use of an assistive device referred to in the CE report?	Yes/No	147	95/52	108/39	81.6%	Moderate
K.1e	Was the ability to dress/undress or other gross/fine hand functions described?	Yes/No	147	54/93	87/60	73.5%	Fair
K.1f	Were weight and height noted?	Yes/No	147	138/9	137/10	98.0%	High
K.1o	Was any part of the physical exam recorded on a standardized form?	Yes/No	147	50/97	54/93	74.1%	Fair
Summary for Section							
All Responses						87.1%	

Source: COMS and SSA Medical Consultant CE Review data.

Notes: The sample includes selected CEs rated jointly by SSA and COMS medical consultants. In total, the COMS and SSA team jointly reviewed 147 adult physical CEs. The rating pairs column indicates how many medical consultants from both teams reviewed the question. The percentage agreement shows the exact agreement between COMS and SSA medical consultants. The summary for "All Responses" includes the percentage agreement for all rating pair responses for every question in the section. The qualitative ratings represent percentage agreement that was "below" threshold (below 70 percent), "fair" (70 to 79 percent), "moderate" (80 to 89 percent), or "high" (above 90 percent) to depict broad trends for each question.

Exhibit VI.6. Physical Exam Findings: Generalist Exams

Question	Response Options	Rating Pairs	SSA	COMS	Percentage Agreement	Qualitative Rating
K.2a Presence or absence of distress?	Yes/No	104	51/53	59/45	73.1%	Fair
K.2b Head, eyes, ears, nose, oral cavity?	At least 1 item/No	104	103/1	102/2	99.0%	High
K.2c Lung auscultation?	Yes/No	104	104/0	103/1	99.0%	High
K.2d Cardiac rhythm?	Yes/No	104	93/11	92/12	93.3%	High
K.2e Cardiac auscultation (heart sounds, murmur, and/or gallop)?	At least 1 item/No	104	100/4	101/3	99.0%	High
K.2f Abdomen, bowel sounds, ascites, tenderness, masses?	At least 1 item/No	104	102/2	100/4	98.1%	High
K.2g Peripheral pulses (wrist or feet) or carotid strength?	Yes/No	104	75/29	80/24	93.3%	High
K.2h Peripheral edema?	Yes/No	104	72/32	75/29	93.3%	High
K.2j Re Joints (including spine) and any myofascial findings?						
K.2j1 Effusion or swelling?	Yes/No	104	49/55	57/47	88.5%	Moderate
K.2j2 Tenderness or trigger/tender points?	Yes/No	104	54/50	62/42	76.9%	Fair
K.2j3 Heat or redness?	Yes/No	104	27/77	35/69	84.6%	Moderate
K.2j4 Synovial thickening?	Yes/No	104	8/96	14/90	92.3%	High
K.2j5 ROM (including spine) in degrees?	Yes/No	104	99/5	100/4	95.2%	High
K2k Muscle bulk or atrophy?	Yes/No	104	46/58	55/49	76.0%	Fair
K.2l Muscle spasm or tone?	Yes/No	104	32/72	50/54	73.1%	Fair
K.2m Straight Leg Raising (SLR)/tension signs in degrees?	Yes, SLR was abnormal	104	12	9	78.8%	Fair
	Yes, SLR was normal		46	58		
	No/Not relevant		46	37		
K.2m(1) If SLR was abnormal, was it confirmed in another body position?	Yes/No	6	1/5	2/4	83.3%	Moderate
K.2n Strength (if abnormal, per specific muscle groups)?	Yes/No	104	92/12	91/13	93.3%	High
K.2o Cranial nerves?	Yes/No	104	56/48	62/42	92.3%	High
K.2p Sensation?	Yes/No	104	86/18	93/11	93.3%	High
K.2q Deep tendon reflexes?	Yes/No	104	93/11	96/8	97.1%	High
K.2r Mental status?	Yes/No	104	46/58	58/46	71.2%	Fair
Summary for Section						
All Responses					88.6%	

Source: COMS and SSA Medical Consultant CE Review data.

Notes: The sample includes selected CEs rated jointly by SSA and COMS medical consultants. In total, the COMS and SSA team jointly reviewed 104 adult physical exam CEs. The rating pairs column indicates how many medical consultants from both teams reviewed the question. The number of rating pairs can be less than 104 if only a subgroup group answered a question. The percentage agreement shows the exact agreement between COMS and SSA medical consultants. The summary for "All Responses" includes the percentage agreement for all rating pair responses for every question in the section. The qualitative ratings represent percentage agreement that was "below" threshold (below 70 percent), "fair" (70 to 79 percent), "moderate" (80 to 89 percent), or "high" (above 90 percent) to depict broad trends for each question.

included details (as per the Green Book) for subspecialty exams, including cardiology, pulmonary, rheumatology, gastro-intestinal, hematology/oncology, endocrinology, renal/genitourinary, and skin diseases. During the initial data extraction, this section included 81 questions to cover these different groups. However, the focus groups and subsequent pretests indicated that such lengthy questions put substantial burdens on the medical consultants.[52] For this reason, the questions were shortened to cover 22 items most central to the types of exams in the Green Book. Additionally, the instructions noted above guided medical consultants to respond to items that were germane to the CE.

All 22 of the questions exceeded the 70 percent threshold.[53] Of the 22 items, 13 had high levels of agreement, including items such as whether or not the exam addressed head, eyes, ears, nose, or oral cavity; lung auscultation; cardiac rhythm; cardiac auscultation, bowel sounds, ascites, abdominal tenderness, and masses; peripheral pulses or carotid strength; peripheral edema; synovial thickening of joints; range of motion in joints in degrees; strength; cranial nerves; sensation; and deep tendon reflexes (K.2b, K.2c, K.2d, K.2e, K.2f, K.2g, K.2h, K.2j4, K.2j5, K.2n, K.2o, K.2p, and K.2q). Three questions had moderate levels of agreement, including whether the exam addressed joint effusion or swelling, heat or redness, and straight leg raise (SLR) and its confirmation (K.2j1, K.2j3, and K.2m). The final six items had fair agreement, including presence or absence of overall claimant distress, joint tenderness or trigger points, muscle bulk or atrophy, muscle spasm or tone, SLR in degrees if SLR was noted, and mental status (K.2a, K.2j2, K.2k, K.2l, K.2m, and K.2r).

c. Physical Exam: Orthopedic/Musculoskeletal

The overall percentage agreement for the orthopedic physical exam was 88 percent (**Exhibit VI.7**). These findings were similar to the physical exam findings above, and likely reflect that most of the items in the musculoskeletal exams were objective (for example, assessment of muscle tone and joint range of motion).

Of the nine questions in this section, eight exceeded the 70 percent threshold. The medical consultants had high agreement on five of the nine questions, including the adequacy of the exam in assessing range of motion, strength, sensation, deep tendon reflexes, and muscle bulk or atrophy (K.3b, K.3d, K.3e, K.3f, and K.3g). An additional question that related to joint stability (K.3h) had moderate agreement. Two questions for addressing muscle spasm and SLR/tension signs had fair agreement (K.3a and K.3c), which, perhaps not surprisingly, was similar to the pattern of agreement noted above for similar items in the generalist exams.[54] The only question that did not meet our 70 percent threshold was whether the SLR/tension signs were confirmed in another body position (63 percent [K.3c1]), though there were only eight rating pairs for this question making it difficult to draw any definitive conclusion about its reliability.

---

[52] Additionally, some of the questions were only asked for very narrow conditions (e.g., speech), which limited the sample size.

[53] One question, K.2m1, had a sample of only six cases, but we have left this information in for completeness.

[54] We did not combine the "yes" responses for K.3c because they addressed different concepts of SLR (normal and abnormal).

Exhibit VI.7. Physical Exams: Orthopedic/Musculoskeletal Exam

Question		Response Options	Rating Pairs	SSA	COMS	Percentage Agreement	Qualitative Rating
K.3a	Did the CE report adequately address muscle spasm or tone?	Yes/No	38	19/19	24/14	76.3%	Fair
K.3b	Did the CE report adequately address joint Range of Motion (ROM) in degrees?	Yes/No	38	37/1	34/4	92.1%	High
K.3c	Did the CE report adequately address SLR/tension signs in degrees?	Yes, SLR was abnormal	38	11	10	78.9%	Fair
		Yes, SLR was normal		21	18		
		No/Not relevant		6	10		
K.3c1	If abnormal, was SLR/tension signs confirmed in another body position?	Yes/No	8	4/4	5/3	62.5%	Below
K.3d	Did the CE report adequately address strength (if abnormal, per specific muscle groups?	Yes/No	38	34/4	32/6	94.7%	High
K.3e	Did the CE report adequately address sensation?	Yes/No	38	31/7	32/6	97.4%	High
K.3f	Did the CE report adequately address deep tendon reflexes?	Yes/No	38	34/4	33/5	92.1%	High
K.3g	Did the CE report adequately address muscle bulk or atrophy?	Yes/No	38	27/11	23/15	89.5%	High
K.3h	Did the CE report adequately address joint instability?	Yes/No	38	11/27	8/30	86.8%	Moderate
Summary for Section							
All Responses						87.8%	

Source: COMS and SSA Medical Consultant CE Review data.

Notes: The sample includes selected CEs rated jointly by SSA and COMS medical consultants. In total, the COMS and SSA team jointly reviewed 38 adult Orthopedic/Musculoskeletal CEs. The rating pairs column indicates how many medical consultants from both teams reviewed the question. The number of rating pairs can be less than 38 if only a subgroup group answered a question. The percentage agreement shows the exact agreement between COMS and SSA medical consultants. The summary for "All Responses" includes the percentage agreement for all rating pair responses for every question in the section. The qualitative ratings represent percentage agreement that was "below" threshold (below 70 percent), "fair" (70 to 79 percent), "moderate" (80 to 89 percent), or "high" (above 90 percent) to depict broad trends for each question.

43

6. Mental Health

There were 142 mental health CEs, and the overall percentage agreement for the mental health section was 93 percent (**Exhibit VI.8**). The objective exam findings were consistent with those noted above for the physical exams.

All 11 of the questions in this section exceeded the 70 percent threshold. There was high agreement on four questions related to procedures about the mental health exam: the claimant's identification, performing the exam via video conference, whether the information was independently elicited, and use of a standardized form (L.1, L.9, L.10, and L.11). Five other questions also had high agreement for assessments, including assessments of speech, thought processes, perceptual abnormalities, mood and affect, and cognition (L.2, L.3, L.5, L.6, and L.7). Two assessment questions for thought content and judgment/insight had moderate agreement (L.4 and L.8).

7. Lab Studies/Exams/Tests

The overall percentage agreement for the lab studies/exams/test ("tests") section was 90 percent (**Exhibit VI.9**). The initial question of the section (M.1) on whether a lab study, exam, or test was ordered applied to all 289 CEs. The remaining questions (M.2–M.5) were only for the 95 CEs in which the pairs agreed that a lab study, exam, or test had been ordered. The strong agreement reflects that the medical consultants could readily identify whether a test had been ordered (in M.1), though there was also moderate to high agreement on four of the other items regarding the test.

All five of the questions in this section exceeded the 70 percent threshold. There was high agreement for whether any tests were ordered (M.1). For the pairs who agreed on the ordering of tests, there was high agreement on whether the tests were in compliance with the Listings of Impairments (M.2). There was also moderate agreement on whether the CE provider discussed the test results in the CE report and whether the worksheet noted that the ancillary study was of a specialized or highly technical nature (M.3 and M.5). Finally, there was fair agreement on whether any lab test or other test associated with the CE Report was unnecessary for adjudication (M.4). Not surprisingly, this last question had lower agreement because it required the medical consultant to make a subjective assessment about the lab test's contribution to case adjudication.

8. CE Report Assessment by the Medical Consultant

The overall percentage agreement for the medical consultants' assessment of the CE report was 82 percent (**Exhibit VI.10**). The overall findings here are of note because the medical consultants had to make several subjective assessments regarding the medical history and objective exams based on the reviews in previous sections. This section covered a broad range of assessments regarding CE processes, content, and quality ratings of earlier CEs (if applicable). Several questions were for all CEs, but some only applied to the initial-level decisions (140 CEs) as opposed to those at the hearings level (149 CEs). Additionally, several questions represented follow-ups to other questions.

.

Exhibit VI.8. Mental Health

Question		Response Options	Rating Pairs	SSA	COMS	Percentage Agreement	Qualitative Rating
L.1	Was there a comment that the claimant's identification was verified during the mental status exam?	Yes/No	142	38/104	35/107	93.7%	High
L.2	Did the CE provider assess: general appearance, behavior, and/or speech?	At least 1 item/No	142	139/3	134/8	96.5%	High
L.3	Did the CE provider assess thought processes?	Yes/No	142	125/17	125/17	90.1%	High
L.4	Did the CE provider assess thought content?	Yes/No	142	122/20	122/20	83.1%	Moderate
L.5	Did the CE provider assess perceptual abnormalities?	Yes/No	142	128/14	126/16	93.0%	High
L.6	Did the CE provider assess mood or affect?	Yes/No	142	133/9	129/13	93.0%	High
L.7	Did the CE provider assess cognition (i.e., concentration, memory, intellectual functioning)?	Yes/No	142	139/3	139/3	97.2%	High
L.8	Did the CE provider assess judgment or insight?	Yes/No	142	121/21	110/32	82.4%	Moderate
L.9	Was the mental status examination independently elicited?	Yes/No	142	139/3	137/5	95.8%	High
L.10	Was the CE performed through a video conference?	Yes/No	142	0/142	0/142	100.0%	High
L.11	Was any part of the Mental Status exam recorded on a standardized form?	Yes/No	142	1/141	1/141	100.0%	High
Summary for Section							
All Responses						93.2%	

Source: COMS and SSA Medical Consultant CE Review data.

Notes: The sample includes selected CEs rated jointly by SSA and COMS medical consultants. In total, the COMS and SSA team jointly reviewed 142 mental health adult CEs. The rating pairs column indicates how many medical consultants from both teams reviewed the question. The percentage agreement shows the exact agreement between COMS and SSA medical consultants. The summary for "All Responses" includes the percentage agreement for all rating pair responses for every question in the section. The qualitative ratings represent percentage agreement that was "below" threshold (below 70 percent), "fair" (70 to 79 percent), "moderate" (80 to 89 percent), or "high" (above 90 percent) to depict broad trends for each question.

Exhibit VI.9. Lab Studies/Exams/Tests

Question	Response Options	Rating Pairs	SSA	COMS	Percentage Agreement	Qualitative Rating
M.1 Were any lab tests, psychological tests, and/or X-rays ordered or added on during the CE?	Yes/No	289	103/186	108/181	92.7%	High
M.2 Were any of the tests not compliant with requirements in the Listings of impairments?	Yes/No	95	3/92	3/92	97.9%	High
M.3 Did the CE provider discuss the test results in the CE report you are reviewing?	Yes/No/ CE provider did not have these results	95	78/13/4	74/19/2	85.3%	Moderate
M.4 Was any lab test or other test, associated with the CE report, unnecessary for adjudication?	Yes/No	95	14/81	25/70	77.9%	Fair
M.5 Did the Worksheet note that the ancillary study needed was of a specialized or highly technical nature?	Yes/No	95	6/89	9/86	88.4%	Moderate
Summary for Section						
All Responses					89.7%	

Source: COMS and SSA Medical Consultant CE Review data.

Notes: The sample includes selected CEs rated jointly by SSA and COMS medical consultants. In total, the COMS and SSA team jointly reviewed 289 adult CEs, which were stratified by type of CE exam (mental versus physical) and determination level (initial versus hearings). The rating pairs column indicates how many medical consultants from both teams reviewed the question. The number of rating pairs can be less than 289 if only a subgroup group answered a question. The percentage agreement shows the exact agreement between COMS and SSA medical consultants. The summary for "All Responses" includes the percentage agreement for all rating pair responses for every question in the section. The qualitative ratings represent percentage agreement that was "below" threshold (below 70 percent), "fair" (70 to 79 percent), "moderate" (80 to 89 percent), or "high" (above 90 percent) to depict broad trends for each question.

Exhibit VI.10. CE Report Assessment by the Medical Consultant

	Question	Response Options	Rating Pairs	SSA	COMS	Percentage Agreement	Qualitative Rating
N.1	Did the CE provider include a discussion of the CE findings?	Yes/No	289	189/100	213/76	77.9%	Fair
N.2	Was a reasonably stated diagnosis provided for each allegation/impairment evaluated by the CE provider?	Yes/No	289	279/10	276/13	94.1%	High
N.3	Were all allegations that SSA intended evaluation of in this CE addressed by the CE provider?	Yes/No	289	250/39	261/28	82.4%	Moderate
N.4	Were all allegations that were evaluated or listed by the provider previously known to SSA?	Yes/No	289	224/65	205/84	66.4%	Below
N.5	Did the CE findings support EVERY diagnosis made by the CE provider?	Yes/No	289	243/46	229/60	75.1%	Fair
N.6	Was a prognosis provided?	Yes/No	289	88/201	82/207	86.9%	Moderate
N.7	Was the prognosis supported by the CE findings?	Yes/No	66	65/1	62/4	92.4%	High
N.8	Were the CE findings and conclusions generally consistent with the MER?	Yes/No	289	263/26	261/28	84.8%	Moderate
N.9	Was there an indication of a change in the applicant's condition that could have affected his/her ability to work?	Yes/No	289	37/252	18/271	83.0%	Moderate
N.10	In your opinion, based on MER at the time the CE was ordered, was the CE needed for adjudication purposes?	Yes/No	140	134/6	137/3	95.0%	High
N.11	Do you agree with the ALJ that the MER was not sufficient to support a claim decision without your current CE?	Yes/No	149	105/44	134/15	65.8%	Below
N.12	Did MER related to the issues evaluated in your CE appear after your CE was performed?	Yes/No	289	92/197	80/209	79.9%	Moderate
N.13	In your opinion, would the late-arriving MER have made your medical consultant CE unnecessary?	Yes/No	57	6/51	11/46	77.2%	Fair
N.14	Were any CEs performed at an earlier adjudicative level in the claim process?	Yes/No	289	75/214	83/206	86.2%	Moderate
N.15	ALJ's stated reason(s) for requesting your CE:						
	Because a new impairment was alleged.	Checked/Not checked	59	2/57	0/59	96.6%	High
	Because of outdated MER or a change in the status of a previously alleged impairment.	Checked/Not checked	59	4/55	3/56	88.1%	Moderate
	Because of a conflict in supporting MER information.	Checked/Not checked	59	2/57	1/58	98.3%	High
	Because a different type of specialty or subspecialty exam was sought to evaluate a previously evaluated allegation.	Checked/Not checked	59	2/57	3/56	91.5%	High
	Because of any other stated reason.	Checked/Not checked	59	2/57	3/56	91.5%	High
	DDS/ALJ did not state any reason for ordering your CE.	Checked/Not	59	48/11	51/8	74.6%	Fair

Question	Response Options	Rating Pairs	SSA	COMS	Percentage Agreement	Qualitative Rating
	checked					
N.16 Was the most recent prior CE from an earlier decision within 6 months of the date the ALJ ordered your CE?	Yes/No	59	12/47	4/55	76.3%	Fair
N.17 If the earlier CE was in your specialty, what was the overall quality of the earlier CE report?	Materially deficient/Average or high quality	59	7/52	4/55	81.4%	Moderate
Summary for Section						
All Responses					81.5%	

Source: COMS and SSA Medical Consultant CE Review data.

Notes: The sample includes selected CEs rated jointly by SSA and COMS medical consultants. In total, the COMS and SSA team jointly reviewed 289 adult CEs, which were stratified by type of CE exam (mental versus physical) and determination level (initial versus hearings). The rating pairs column indicates how many medical consultants from both teams reviewed the question. The number of rating pairs can be less than 289 if only a subgroup group answered a question. The percentage agreement shows the exact agreement between COMS and SSA medical consultants. The summary for "All Responses" includes the percentage agreement for all rating pair responses for every question in the section. The qualitative ratings represent percentage agreement that was "below" threshold (below 70 percent), "fair" (70 to 79 percent), "moderate" (80 to 89 percent), or "high" (above 90 percent) to depict broad trends for each question.

Of the 22 questions in this section (which included 16 binary questions and one question with six response options [N.15]), 20 exceeded the 70 percent threshold. The first seven questions in the section provided summary information on the reviewed characteristics for all CEs, including a discussion of findings by the CE provider, description of allegations, and prognosis (N.1-N.7). For these questions, there was high agreement regarding whether there was a reasonably stated diagnosis and a prognosis supported by CE findings (N.2 and N.7); moderate agreement for two items related to case allegations and CE provider's prognosis (N.3 and N.6); and fair agreement for two items considering whether the CE report included a discussion of the CE's findings and whether the CE findings supported every diagnosis made by the CE provider (N.1 and N.5). Agreement on whether the allegations evaluated or listed by the provider were previously known to SSA fell below the 70 percent threshold (N.4).

The next six questions (N.8–N.13) related to MER. For the 140 CEs at the initial level, there was high agreement on whether the CE was needed for adjudication purposes based on the MER (N.10); moderate agreement for four items related to consistency between the CE findings and the MER, any indications of change in the applicant's allegations, and whether the MER potentially related to the CE was received after the CE was performed (N.8, N.9, and N.12); and fair agreement on whether the late-arriving MER would have made the medical consultant CE unnecessary (N.13). The latter item was a follow-up for 57 CEs to the N.12 question. The percentage agreement for the question regarding whether the medical consultant agreed with the ALJ that the MER was not sufficient to decide the appeal (N.11) was below 70 percent.

The final five questions (N.14–N.18) narrowed the CEs down to a subset of 59 CEs in which there was a CE ordered at an earlier adjudication level. First, N.14 asked if any CE's were performed at an earlier adjudicative level in the case process, which had moderate agreement. The six-part N.15 question on the ALJ's stated reasons for requesting the CE varied in agreement, ranging from fair agreement (75 percent) on when the ALJ did not state any reason for ordering the CE to high agreement (98 percent) on when the CE was ordered because of a conflict in MER information. The final two questions assessing the quality rating of the earlier CE and whether the ALJ waited six months before ordering a CE had moderate agreement and fair agreement, respectively.

In summary, the only two questions that did not meet reliability thresholds in this section were based on subjective judgments that might have related in part to difficulty extracting information in the folder. Question N.4 asked medical consultants what SSA *might* have known about previous allegations. Question N.11 asked medical consultants to make a "subjective call" on whether they agreed with the ALJ that the CE should have been ordered based on the available MER. The remaining questions in this section were all of at least fair reliability.

9. Medical Source Statement and Functional Capacities

The overall percentage agreement for the MSS and Functional Capabilities section was 83 percent (**Exhibit VI.11**). There was one question for all CEs, and the remaining questions were split between capabilities assessments for the 147 physical CEs and the 142 mental CEs. Although a standardized MSS form existed, several CEs did not include it. For example, COMS medical consultants found that 156 of the 289 CEs had a formal MSS on file (see O.1). Without a formal MSS, the medical consultants had to search for this information in other areas.

Exhibit VI.11. Medical Source Statement and Functional Capacities

Question		Response Options	Rating Pairs	SSA	COMS	Percentage Agreement	Qualitative Rating
O.1	Was a formal medical source statement (MSS) on file?	Yes/No	289	156/133	153/136	80.3%	Moderate
O.2	Which of the following functional capacities were estimated for an adult physical CE?						
O.2a	Sitting (for how long?).	Yes/No	147	86/61	90/57	90.5%	High
O.2b	Standing (for how long?).	Yes/No	147	87/60	89/58	90.5%	High
O.2c	Walking (for how long or how far?).	Yes/No	147	87/60	90/57	88.4%	Moderate
O.2d	Lifting (how much?).	Yes/No	147	88/59	91/56	88.4%	Moderate
O.2e	Carrying (how much?).	Yes/No	147	84/63	89/58	88.4%	Moderate
O.2f	Handle/finger objects.	Yes/No	147	89/58	90/57	85.7%	Moderate
O.2g	Hearing.	Yes/No	147	69/78	83/64	81.0%	Moderate
O.2h	Speaking.	Yes/No	147	46/101	71/76	73.5%	Fair
O.2i	Travel.	Yes/No	147	38/109	61/86	76.2%	Fair
O.3	Which of the following functional capacities were estimated for an adult Mental Health CE?						
O.3a	Understanding and memory.	Yes/No	142	113/29	125/17	81.7%	Moderate
O.3b	Concentration, persistence, and pace.	Yes/No	142	102/40	118/24	70.4%	Fair
O.3c	Social functioning.	Yes/No	142	114/28	119/23	81.0%	Moderate
O.3d	Adaptation.	Yes/No	142	95/47	100/42	82.4%	Moderate
O.3e	Capability of handling funds.	Yes/No	142	119/23	130/12	83.8%	Moderate
Summary for Section							
All Responses						82.7%	

Source: COMS and SSA Medical Consultant CE Review data.

Notes: The sample includes selected CEs rated jointly by SSA and COMS medical consultants. In total, the COMS and SSA team jointly reviewed 289 adult CEs, which were stratified by type of CE exam (mental versus physical) and determination level (initial versus hearings). The rating pairs column indicates how many medical consultants from both teams reviewed the question. The number of rating pairs can be less than 289 if only a subgroup group answered a question. The percentage agreement shows the exact agreement between COMS and SSA medical consultants. The summary for "All Responses" includes the percentage agreement for all rating pair responses for every question in the section. The qualitative ratings represent percentage agreement that was "below" threshold (below 70 percent), "fair" (70 to 79 percent), "moderate" (80 to 89 percent), or "high" (above 90 percent) to depict broad trends for each question.

All 15 questions in this section exceeded the 70 percent threshold. For the overall question of whether there was a formal MSS on file for all 289 CEs, there was moderate agreement. For the 147 physical CEs, there was high agreement on two questions for sitting and standing (O.2a and O.2b); moderate agreement for five questions related to walking, lifting, carrying, handling objects, and hearing (O.2c, O.2d, O.2e, O.2f, and O.2g); and fair agreement for two questions related to speaking and travelling (O.2h and O.2i). For the 142 mental health CEs, there was moderate agreement for four questions related to understanding/memory, social function, adaption, and capability of handling funds (O.3a, O.3c, O.3d, and O.3e); and fair agreement for concentration, persistence, and pace (O.3b).

10. Overall Completeness of CE Report

The overall percentage agreement for the section on completeness of the report was 85 percent (**Exhibit VI.12**). This section included one objective question about whether the CE was signed by a CE provider (P.1). The remaining two questions (P.2 and P.3) required qualitative assessments by the medical consultants regarding whether the CE was of strong quality and whether SSA had obtained what it had ordered. To achieve acceptable quality for these questions, we collapsed the categories of the two review items. For the overall quality of the report (P.2), we distinguished between reports that were materially deficient, which meant they contained a critical error and could not be used to adjudicate the case properly, and all other quality ratings (including average and high quality, which both imply they could be used to adjudicate the case). For the completeness question (P.3), we identified medical consultants who either disagreed or strongly disagreed that the CE report contained all of the information SSA had paid for. We tried variations of other response category options for questions P.2 and P.3, though the two presented here offered the strongest reliability ratings, which drove up the agreement of the questions in this section.

All three of the questions in this section exceeded the 70 percent threshold. There was high agreement regarding whether the report was signed by an acceptable medical source (P.1) and moderate agreement for the medical consultants' assessment of the CE's quality and completeness (P.2 and P.3). The findings for these questions will be especially useful in making assessments of CE quality for our next report.

B. Examiner Sections

Below, we review the findings from the 25 CEs jointly reviewed by the examiners, illustrated in two exhibits. The examiners' instrument had six sections, but the number of questions in each section was fairly limited, so we summarize the findings in two exhibits. The first exhibit includes the primary data elements used by the design team to assess the triage process, such as the type of exam. The second exhibit includes other items extracted by the examiner that provide descriptive information on CE processes, such as the follow-up with CE provider.

1. Type of CE, Case Dates, and Qualifications of CE Provider

The overall percentage agreement for Type of CE, Case Dates, and Qualifications of the Provider was 98 percent (**Exhibit VI.13**). These high levels of agreement confirmed that the examiner was using the appropriate methods to identify CE characteristics, particularly for the questions B.1 and B.2 that were most relevant to the triage process. This finding gave the design team confidence that the COMs and SSA management teams could reliably triage the CEs to the appropriate medical consultant.

Exhibit VI.12. Overall Completeness of CE Report

Question		Response Options	Rating Pairs	SSA	COMS	Percentage Agreement	Qualitative Rating
P.1	Was the CE report signed by an acceptable medical source (provider) who actually performed the CE?	Yes/No	289	263/26	278/11	90.7%	High
P.2	What is the overall quality of the CE report you are primarily reviewing?	Materially deficient/Average or high quality	289	35/254	35/254	82.7%	Moderate
P.3	The CE report contained all of the information (findings, conclusions, responses to questions) that SSA paid for.	Strongly agree, agree/Neither agree nor disagree/Disagree or strongly disagree	289	239/50	244/45	82.4%	Moderate
Summary for Section							
All Responses						85.3%	

Source: COMS and SSA Medical Consultant CE Review data.

Notes: The sample includes selected CEs rated jointly by SSA and COMS medical consultants. In total, the COMS and SSA team jointly reviewed 289 adult CEs, which were stratified by type of CE exam (mental versus physical) and determination level (initial versus hearings). The rating pairs column indicates how many medical consultants from both teams reviewed the question. The number of rating pairs can be less than 289 if only a subgroup group answered a question. The percentage agreement shows the exact agreement between COMS and SSA medical consultants. The summary for "All Responses" includes the percentage agreement for all rating pair responses for every question in the section. The qualitative ratings represent percentage agreement that was "below" threshold (below 70 percent), "fair" (70 to 79 percent), "moderate" (80 to 89 percent), or "high" (above 90 percent) to depict broad trends for each question.

Exhibit VI.13. Type of CE, Case Dates, and Qualifications of the Provider

Question	Response Options	Rating Pairs	SSA	COMS	Percentage Agreement	Qualitative Rating
Type of CE						
B.1 Type of exam	Adult physical/ Adult mental/ Child physical/ Child mental	25	12/13/0/0	12/13/0/0	100%	High
B.2 For adult physical only, what type of exam was received?	16 categories	25	16[a]	16[a]	100%	High
B.3 For adult mental only, what type of exam was received?	Mental Status examination/ Psychological Testing/Both	13	10/2/1	8/0/5	69.2%	Below
Case Dates						
E.1 What was the date the CE was requested by the DDS or ALJ?	Known date/Unknown	25	22/3	23/2	96.0%	High
E.3 What was the date the CE report was received by the DDS or ALJ?	Known date/Unknown	25	25/0	25/0	--[b]	
(Calculated) Days between request and receipt	30 days or over/Under 30 days	22	12/10	12/10	100%	High
Qualifications of Provider						
F.1 What was the licensure (profession) of the CE provider?	Licensed physician/ Licensed psychologist/Other	25	12/13/0	12/13/0	100%	High
F.2 Was the CE provider's license status noted (must show expiration date in CE report)?	Yes/No	25	24/1	24/1	100%	High
F.3 What was the CE provider's name? (MD/DO's only)?	Name	25	25[c]	25[c]	100%	High
F.4 In what State was the CE performed?	State	25	25[c]	25[c]	100%	High
F.5 Was the CE provider a treating source?	Yes/No	25	0/25	0/25	100%	High
F.6 Was a treating source asked to perform the CE?	Yes/No or unknown	25	0/25	0/25	100%	High
Summary for Section						
All Responses					98.1%	

Source: COMS and SSA Examiner CE Review data.

Notes: The sample includes selected CEs rated jointly by SSA and COMS examiner. In total, the COMS and SSA team jointly reviewed 25 adult CEs, which were stratified by determination level (initial versus hearings). The rating pairs column indicates how many examiners from both teams reviewed the question. The summary for "All Responses" includes the percentage agreement for all rating pair responses for every question in the section. The qualitative ratings represent percentage agreement that was "below" threshold (below 70 percent), "fair" (70 to 79 percent), "moderate" (80 to 89 percent), or "high" (above 90 percent) to depict broad trends for each question. COMS and SSA examiners matched for all 16 categories for B2. [b] COMS and SSA examiners reported dates of scheduling. For this question, we use the average days between the request and received date to assess agreement for medical consultant pairs because the time frame between these two dates is of analytic interest. [c] Names of CE provider and states all perfectly match.

Ten of the 11 questions in these combined sections exceeded the 70 percent threshold. Notably, Question B.1 on the type of exam (adult mental or adult physical) showed 100 percentage agreement). There was also 100 percent agreement for the type of physical exam (B.2) (internal medicine, musculoskeletal, or neurology), which was important because several items in the instrument depended on this identification. On the third question (B.3), which applied to adult mental CEs, the rate of agreement was below 70 percent. Disagreement on this question likely occurred when one examiner recorded testing as part of the mental status exam and the other recorded testing as separate. Although this question did not meet the reliability threshold, it did not have any impact on the triage because this information was not needed to triage mental health CEs by COMS's and SSA's management teams. The remaining eight questions on case dates and qualifications of CE providers had high agreement, including 100 percent agreement on seven of them (E.1, E.3, F.1, F.2, F.3, F.4, F.5, and F.6).

2. Other Processes (Process of Obtaining a CE, Medical Source Statement from CE Provider, and Follow-Up with Provider)

The overall percentage agreement for other CE processes was 64 percent, though this agreement is understated, because many of the questions had continuous response categories (**Exhibit VI.14**). Nonetheless, several questions did not meet reliability standards. The findings here suggest that the information in the other sections, particularly regarding the information on medical sources and the specific DDS or ALJ requests for an MSS, was potentially harder to find and/or more subjective. One potential challenge in locating this information was that the authorization letter for the CE was not in the administrative folder.[55] This lower level of agreement might also reflect the fact that the examiners did not receive formal guidance for these questions.

Of the nine questions in these three examiner sections, only four met our percentage agreement threshold. For the number of medical sources identified and fees (D.1. and D.6), there was below 70 percent agreement, though we find these variables as meeting our thresholds given the correlation rates for both of them were above 90 percent. However, three questions did not meet reliability thresholds, including: (1) how many medical sources provided information, (2) the number of medical sources that provided information before the CE was initiated, and (3) how long the examiner waited after the last request for MER before ordering the CE (D.2, D.3, and D.4). Both questions about the procedures for requesting an MSS, including whether it was expected (in the worksheet or in the ALJ's opinion) and whether it was in the CE authorization or invoice, were below 70 percentage agreement (H.1 and H6). Lastly, the two questions with high agreement were the final question about follow-up contact with the CE provider (Q.1) as well as the fifth question of whether a DDS medical consultant either requested the CE or agreed with the examiner's decision to order a CE (D.5).

[55] Specifically, some information needed for CEs is placed in the current development section of the electronic folder. At every decision point for the case, once a decision is made the current development section is deleted at the State DDS level.

Exhibit VI.14. Process of Obtaining a CE, Medical Source Statement from CE Provider, and Follow- Up with Provider

	Response Options	Rating Pairs	SSA	COMS	Percentage Agreement	Correlation Coefficient	Qualitative Rating
Process of Obtaining a CE- Initial and Hearings							
D.1 How many Medical Sources (MSs) were identified on the 3368 or 3820?	Number	12	3.9	4.3	66.6%	0.99	Fair[a]
D.2 How many MSs provided medical information? (There may be more in the file than are listed on the 3368/3820.)	Number	12	3.5	6.7	33.3%	0.54	Below
D.3 Number of MSs providing medical information before initiating the CE?	Number	12	3.2	6.3	33.3%	0.53	Below
D.4 How long did the disability examiner wait (after the last request for MER) before purchasing the CE?	Less than 21 days or 1 month/More than 1 month	12	2/10	7/5	58.3%	NA	Below
D.5 Did a DDS medical consultant either request/agree with the examiner's decision?	Yes/No or unknown	12	1/12	0/12	91.6%	NA	High
D.6 What was the cost of the basic ("hands-on") CE? (For mental health CEs that include cognitive testing, include the cost[s] of the tests)	Dollar amount (if known)/Unknown	25	$149.62/ 12	$188.10/ 10	72.0%	0.92	Fair[a]
Medical Source Statement from CE Provider							
H.1 Did the DDS Worksheet or ALJ's opinion note that an MSS was expected/requested?	Yes/No	25	2/23	9/16	56.0%		Below
H.2 Did the CE authorization or invoice request an MSS?	Yes/No	25	8/17	16/9	52.0%	NA	Below
Follow- Up with CE Provider							
Q.1 Was there any follow-up contact with the CE provider?	Yes/No or unknown	25	1/24	0/25	96.0%	NA	High
Summary for Section							
All Responses					64.3%		

Source: COMS and SSA Examiner CE Review data.

Notes: The sample includes selected CEs rated jointly by SSA and COMS examiner. In total, the COMS and SSA team jointly reviewed 25 adult CEs, which were stratified by determination level (initial versus hearings). The rating pairs column indicates how many examiners from both teams reviewed the question. The summary for "All Responses" includes the percentage agreement for all rating pair responses for every question in the section. The qualitative ratings represent percentage agreement that was "below" threshold (below 70 percent), "fair" (70 to 79 percent), "moderate" (80 to 89 percent), or "high" (above 90 percent) to depict broad trends for each question.

[a]The qualitative rating of "fair" applies to variable based on correlation coefficient above 0.90.

VII. DISCUSSION

The data collected for this study represent an important step in documenting information about CE processes, content, and quality that was not available in any other data source. Given that CEs were part of more than 40 percent of claims in 2009, it is important for SSA to track whether the information it is receiving meets expectations. The findings in this study indicate that it is possible to train multiple medical consultants to collect consistent and detailed information on the quality of CEs despite the subjectivity inherent in this task.

The IRR analysis in this study played an important role in ensuring the quality of data interpretation. Using a multi-step process, we were able make adjustments and refine the data collection instrument throughout this study to better achieve SSA's goals. The IRR analysis of the initial data extraction that started in 2010 showed the percentage agreement rates for several questions were below 70 percent for a sample of 129 matched CEs. In response, the design team significantly revised the instrument, created a codebook to inform data extraction, and conducted a training session with COMS and SSA medical consultants. During this revision process, the members of the COMS medical consultant review team tested multiple versions using small-scale IRR analyses before the full test with SSA medical consultants. This persistent testing of the instrument led to a substantially improved instrument that could have been used to finish the full data extraction for the study were it not for time pressures to cut the sampling short.

Our findings indicate that most questions met or substantially exceeded the reliability thresholds set for the instruments. The overall percentage agreement across all sections was 85 percent, well above the 70 percent threshold established for the study (**Exhibit V.1**). Of the 142 questions in the instrument, 128 questions exceeded the reliability thresholds, including 99 that were above moderate agreement (i.e., above 80 percent agreement). Across sections, the percentage agreement ranged from 78 percent (Medical History Section) to 95 percent (Type of CE, Case Dates, and Qualifications of the Providers).

Our findings should be viewed as an important step to assist SSA in meeting the long-term CE monitoring objectives. Nonetheless, the high degree of reliability indicates the promise of using the instruments to extract data for much larger samples. Of particular importance for future studies is to assess the extent to which state differences affect CE processes, content, and quality.

With the transfer of the web-based data extraction tool to SSA, there are several possible options for using the findings in this report to expand CE Review data extraction efforts. One option is to recruit internal SSA examiners and/or medical consultants, similar to the ones used in this study, to expand the sample of CEs reviewed. This option could be used to address research questions that could not be fully examined under this project, such as whether substantial differences exist in CE content or quality across states or CE providers. A related option is to use the data extraction tool for SSA's operational purposes and track CE processes over time and across states for key indicators. For example, examiners or medical consultants could review CEs on an ongoing basis, and the findings from select indicators could feed into CE management reports to assess whether CEs are ordered in accordance with regulations.

Another option for SSA is to explore using examiners as reviewers for the medical consultant sections, which would cut the costs of the review process. Under this option, SSA could use the IRR data collected in this study to test whether examiners could achieve the same reliability standards as those found in this report by completing the medical consultant sections of the instrument using the

same CEs used in this study. Such a change would significantly reduce the costs of collecting data on CEs on a regular basis.

Regardless of whether an examiner, medical consultant or other staff member conducts the review, the findings from the study underscore the importance of having all future reviewers go through a brief training, carefully review the codebook for the medical consultant sections, and to test the reliability of the evaluations before extracting data for a much larger sample. This type of upfront preparation should equip future reviewers to extract data in a reliable manner in accordance with the thresholds established for this report.

Finally, the findings here confirm that the instrument developed to address SSA's key research questions for the project, especially those related to §404.1519t and §416.919t., are reliable. The second report for this project will provide a detailed discussion of the questions and a summary of the policy implications. In that report, we will include the findings presented here, to underscore the strength of the instrument and the credibility of this research effort which, although it employed a small sample, is the only current study of CE processes, content, and quality.

REFERENCES

Hartling, L., M. Hamm, A. Milne, B. Vandermeer, P.L. Santaguida, M. Ansari, A. Tsertsvadze, S. Hempel, P. Shekelle, and D.M. Dryden. "Validity and Inter-Rater Reliability Testing of Quality Assessment Instruments." AHRQ Publication No. 12-EHC039-EF. Rockville, MD: Agency for Healthcare Research and Quality, March 2012.

Institute of Medicine. Committee on Improving the Disability Decision Process: SSA's Listing of Impairments and Agency Access to Medical Expertise. "Improving the Social Security Disability Process: Interim Report." Washington, DC: National Academies of Science, 2006.

Landis, J.R. "The Measurement of Observer Agreement for Categorical Data." vol. 33, 1977, pp. 159-174.

Perez-Johnson, Irma, Ken Fortson, Christine Ross, Claudia Gentile, Samia Amin, Hanley Chiang, and Larissa Campuzano. "Design Considerations for a Study to Validate Measures of Teacher Classroom Practices." Princeton, NJ: Mathematica Policy Research, 2009.

Raudenbush, Stephen W. and Sally Sadoff. "Statistical Inference when Classroom Quality is Measured with Error." *Journal of Research on Educational Effectiveness,* vol. 1, no. 2, 2008, pp. 138-154.

Social Security Advisory Board. "Aspects of Disability Decision Making: Data and Materials." Washington, DC: Social Security Advisory Board, 2012.

Stemler, Steven E. "A Comparison of Consensus, Consistency, and Measurement Approaches to Estimating Interrater Reliability." *Practical Assessment, Research & Evaluation,* vol. 9, 2004, pp. 1-1.