Assessing the Utility of the Work Disability Functional Assessment Battery (WD-FAB) in the Continuing Disability Review Process

A Final Report on the WD-FAB Research Study

NIH Agreement Number: XCL23001-002
SSA Agreement Number: NIH-24-0001

June 30, 2024

TABLE OF CONTENTS

LIST OF TABLES **3**
LIST OF FIGURES **4**
ACRONYMS AND ABBREVIATIONS **5**
DEFINITIONS **6**
EXECUTIVE SUMMARY **7**
INTRODUCTION **9**
THE WORK DISABILITY FUNCTIONAL ASSESSMENT BATTERY (WD-FAB) **10**
THE WD-FAB RESEARCH STUDY **11**
UTILITY OF THE WD-FAB IN THE CDR PROCESS **13**
SINGLE WD-FAB ADMINISTRATION 14
CHANGE IN WD-FAB DATA 26
VALUE AND FEASIBILITY OF INTEGRATING THE WD-FAB INTO THE CDR PROCESS 41
SUMMARY RECOMMENDATIONS **47**
CONCLUSION **49**
REFERENCES **51**
APPENDIX A **52**
APPENDIX B **54**
APPENDIX C **60**
APPENDIX D **62**

LIST OF TABLES

Table 1. SSA Administrative Data for Analytic Sample 12
Table 2. Mean WD-FAB scores by CDR Diary Date Category 17
Table 3. Mean WD-FAB scores by Predictive Model Score Category 18
Table 4. CDR Mailer (SSA-455) Questions and Response Options 19
Table 5. Logistic Regression Comparing WD-FAB to CDR Mailer (Survey 1) 19
Table 6. Logistic Regression Comparing WD-FAB to CDR Mailer (Survey 2) 20
Table 7. Mean Time from WD-FAB Survey Administration to Decision Date in Days 24
Table 8. Distribution and Continuance Rate for WD-FAB Profiles by Predictive Model Score 25
Table 9. Logistic Regression Comparing WD-FAB to CDR Decision 26
Table 10. Detectable Change in WD-FAB Scale Scores (90% Confidence Interval) 27
Table 11. Number of Detectable Scale Score Changes (90% Confidence Interval) 28
Table 12. Number of Detectable Changes by Impairment Type (90% Confidence Interval) 28
Table 13. Demographics by Category of Change in Function 31
Table 14. Logistic Regression Comparing Change in Function to Predictive Model Scores 33
Table 15. Logistic Regression Comparing Change in WD-FAB to CDR Mailer (Survey 2) 34
Table 16. Odds Ratios Comparing Change in WD-FAB with Change in CDR Mailer 35
Table 17. Change in Case Presence from CPD to CDR Compared with WD-FAB Change 36
Table 18. Presence of Functioning Information as Average Mentions per Sentence and per Page by WD-FAB Change in Function Category 37
Table 19. CDR Initial Allowance Rate by Change Sub-Group 37
Table 20. Administration Metrics by Survey Time Point and Mode 41
Table 21. Focus Group Feedback Categories, Themes, and Specific Examples 45

LIST OF FIGURES

Figure 1. Analytic Sample Demographics by Administration Mode 13
Figure 2. Box Plot of WD-FAB Scores at Survey 1 (S1) and Survey 2 (S2) 14
Figure 3. Distribution of WD-FAB Profiles at Survey 1 and Survey 2 15
Figure 4. Distribution of WD-FAB Profiles by SSA Sub-Population (Survey 1) 16
Figure 5. Distribution of WD-FAB Profiles by SSA Sub-Population (Survey 2) 16
Figure 6. Distribution of WD-FAB Profiles by CDR Diary Type (Survey 1 and 2) 17
Figure 7. Distribution of WD-FAB Profiles by Predictive Model Score (Surveys 1 and 2) 18
Figure 8. Box Plot of WD-FAB Basic Mobility Scores by Mean Reported Difficulty in SSA-454 Daily Activities 21
Figure 9. Number of CDR cases with Functioning Information Across WD-FAB scales by WD-FAB Functional Level 22
Figure 10. Proportion of Cases with Functioning Information by WD-FAB Functional Level 23
Figure 11. Mean Mention Count by WD-FAB Functional Level 23
Figure 12. Box Plot of WD-FAB Scores for Continuances (CO) vs. Cessations (CE) at Survey 1 24
Figure 13. Box Plot of WD-FAB Scores for Continuances (CO) vs. Cessations (CE) at Survey 2 25
Figure 14. Scatterplot of Standardized Differences Between WD-FAB Administrations 27
Figure 15. Log Odds of Improvement Beyond 90% Confidence Interval Threshold for Change 29
Figure 16. Receiver Operator and Precision-Recall Curves for Model Predicting Improvement 30
Figure 17. Log Odds of Decline Beyond 90% Confidence Interval Threshold for Change 30
Figure 18. Receiver Operator and Precision-Recall Curves for Model Predicting Decline 31
Figure 19. Distribution of Change in Function by CDR Diary Type (Scores vs. Profiles) 32
Figure 20. Distribution of Change in Function by Predictive Model Score (Scores vs. Profiles) 33
Figure 21. Log Odds for Predicting Cessation Cases (Mean and 95% Confidence Interval) 38
Figure 22. Receiver Operator and Precision-Recall Curves for Model Predicting Cessation 39
Figure 23. Presence of Functioning Information in Medical Evidence for Continuances vs. Cessations .. 39
Figure 24. Ratios of Presence of Functioning Information in the Disability Determination Explanation for Cessations over Continuances 40
Figure 25. Pie chart of Likert-type Responses for whether Questions were Burdensome 42
Figure 26. Pie chart of Likert-type Responses for whether Questions were Confusing 43
Figure 27. Pie chart of Likert-type Responses for whether Questions were Difficult 43
Figure 28. Pie chart of Likert-type Responses for whether Comfortable Answering Questions 44

Acronyms and Abbreviations

ADL:	Activity of Daily Living
AUPRC:	Area under the precision recall curve
AUROC:	Area under the receiver operator characteristic curve
BM:	Basic Mobility
BS:	Body System
CAT:	Computer Adaptive Testing
CC:	Communication & Cognition
CDR:	Continuing Disability Review
CE:	Cessation
CM:	Community Mobility
CO:	Continuance
CPD:	Comparison Point Decision
DCN:	Decision
DDE:	Disability Determination Explanation
FMF:	Fine Motor Function
FMR:	Full Medical Review
FY:	Fiscal Year
IRT:	Item Response Theory
ME:	Mood & Emotions
MIE:	Medical Improvement Expected
MINE:	Medical Improvement Not Expected
MIP:	Medical Improvement Possible
MIRS:	Medical Improvement Review Standard
NIH:	National Institutes of Health
NLP:	Natural Language Processing
PMS:	Predictive Model Scores
RS:	Resilience & Sociability
S1:	Survey 1
S2:	Survey 2
SD:	Standard Deviation
SED:	Supported Employment Demonstration
SGA:	Substantial Gainful Activity
SR:	Self-Regulation
SSA:	Social Security Administration
SSDI:	Social Security Disability Insurance
SSI:	Supplemental Security Income
UBF:	Upper Body Function
WD-FAB:	Work Disability Functional Assessment Battery

Definitions

WD-FAB Functional Levels

Above Average:	Minimal or no limitation
Average:	Somewhat limited
High:	Some or mild limitations
Highest:	No limitation
Low:	Periodic significant limitations
Lowest:	Persistent and significant limitations

WD-FAB Functional Profiles

Profile 1 (P1):	2 or more Lowest/Low WD-FAB scores
Profile 2 (P2):	Exactly 1 Lowest/Low WD-FAB score
Profile 3 (P3):	2 or more Average and no Lowest/Low WD-FAB scores
Profile 4 (P4):	Exactly 1 Average and no Lowest/Low WD-FAB scores
Profile 5 (P5):	All WD-FAB scores greater than Average

Executive Summary

This report provides recommendations for three potential applications of the Work Disability Functional Assessment Battery (WD-FAB) in the SSA continuing disability review (CDR) process for SSA to consider evaluating in future studies. These recommendations are based on analyses the NIH conducted using data SSA collected as part of the WD-FAB Research Study as well as through SSA's regular business processes for adjudicating medical CDRs.

Application 1. Incorporate WD-FAB scores into the CDR Predictive Models

The WD-FAB can be used as part of the screening process to help SSA determine who should undergo a full medical review (FMR). From these analyses, we see associations between baseline WD-FAB scale scores (Upper Body Function, Fine Motor Function, Mood & Emotions) as well as changes in scores (Upper Body Function, Mood & Emotions) and the likelihood of cessation such that including WD-FAB data as part of the current predictive models has the potential to refine the model estimates further. SSA currently runs the predictive models annually for all beneficiaries and then looks at scores for those beneficiaries whose CDR diaries have come due. Therefore, in order to be able to incorporate WD-FAB data into the predictive models, we recommend administering the WD-FAB to all beneficiaries annually. With an annual data collection, both the baseline WD-FAB scores as well as the year-over-year change in scores could be used as factors in the predictive models. This would provide SSA with more regular insight into beneficiaries' functional status and how their function changes over time. Having access to such data also offers an opportunity to potentially refine the diary types and how diary dates are assigned in the future.

Application 2. Use WD-FAB scores in addition to or instead of the CDR Mailer

The WD-FAB can be collected as part of, or instead of, Form SSA-455 (the CDR Mailer). When comparing WD-FAB scores to the CDR mailer responses, we found relationships between both the physical and mental functioning scales and likelihood of having been employed in the past two years, health status, and healthcare use. The relationships between the WD-FAB scores for Upper Body Function and Resilience & Sociability and employment status are consistent with results from the Supported Employment Demonstration, which found similar relationships over a three-year period. In addition to supporting the existing CDR mailer questions, the WD-FAB has the added advantages of providing data that are more consistent, easier to compare across respondents and administrations, and more comprehensively measure functional abilities that are related to work. Leveraging WD-FAB data promotes evidence-based decision making that can look beyond just the completeness of the CDR mailer responses to consider both current level of functioning and whether a beneficiary has demonstrated sufficient change in function to be likely to earn above SGA and therefore warrant spending resources to conduct a FMR.

Application 3. Assess function as part of the Full Medical Review

WD-FAB data can serve multiple purposes as part of the full medical review. The WD-FAB can be used to supplement existing processes such as refining or replacing the daily activities section of Form SSA-454 (the CDR Report). All eight WD-FAB scales demonstrated expected relationships with the current SSA-454 Daily Activities with lower WD-FAB scores, which represents lower functional abilities, corresponding to greater reported difficulties with the daily activities on average. The WD-FAB has the added advantages of covering a wider range of activities and providing greater granularity beyond the binary 'Yes' or 'No' of the current form.

The WD-FAB could be leveraged to assist with case development by identifying areas of limitation that should be investigated and developed within the medical evidence. By comparing with the functioning information from the medical records, we showed that WD-FAB data can be associated with the medical evidence such that decisions can still be made on the basis of objective findings. We observed similar patterns in the amount of functioning information available in the medical records and used to support determinations across existing CDR measures and the WD-FAB suggesting that the WD-FAB can be just as relevant as other tools used within the CDR process while having the advantage of rigorous scientific validation and development specifically for SSA's use and client population. Since the FMR is driven by the medical improvement review standard, WD-FAB data can also be used to measure functional change as part of evaluating signs and symptoms from the comparison point decision (CPD). In particular, the WD-FAB offers a method for collecting comparable information on functioning across CPD and CDR cases that will be more consistent than relying on availability of such information from the medical records. However, under current SSA business regulations, in order for WD-FAB data to be an admissible measure to be included as part of the side-by-side comparison between CPD and CDR evidence, the WD-FAB would have to be administered within the CPD.

The findings from the WD-FAB Research Study indicate that the WD-FAB, as a measure of functional change over time, has a similar relationship to evidence cited for CDR decision making as existing CDR measures while having the added advantage of being more standardized, comprehensive, and easier to compare over multiple administrations. This study provides a foundation of evidence for assessing the application and utility of the WD-FAB as part of SSA's CDR business processes. Additional work is needed to sufficiently justify and support the integration of the WD-FAB as a complementary information source. This includes the refinement and development of resources to help adjudicators understand and interpret WD-FAB data in the SSA context, as well as studies that include the WD-FAB as part of actual CDR cases to assess impact on decision making. With constrained resources available for conducting FMRs, SSA will need to seek new approaches for maximizing the effectiveness of the CDR program. These preliminary findings demonstrate that there is potential value and benefit for SSA in pursuing these next steps and moving towards use of the WD-FAB in the CDR process.

Introduction

The Social Security Administration (SSA) is required by law to conduct continuing disability reviews (CDRs) to determine Social Security Disability Insurance (SSDI) and Supplemental Security Income (SSI) beneficiaries' continuing program eligibility [1]. CDRs are conducted as part of SSA's program integrity efforts to ensure that only those beneficiaries who remain disabled receive benefits. The CDR process tends to yield favorable savings-to-cost ratios with an average of $14.1 to $1 reported in fiscal year (FY) 2014, although this ratio varies depending on the types of cases reviewed and whether sufficient resources are available to review all CDR cases due in a given year [2]. There are two types of CDRs – work CDRs and medical CDRs. A work CDR can be initiated when SSA has information that a beneficiary has earnings above substantial gainful activity (SGA) levels and is not participating in the Ticket to Work program. Medical CDRs are scheduled based on how likely an impairment is to improve, indicated by the CDR diary date. The three types of CDR diaries are Medical Improvement Expected (MIE), reviewed every 6 to 18 months; Medical Improvement Not Expected (MINE), reviewed every 7 years; and Medical Improvement Possible (MIP), reviewed every 3 years [3]. The diary date is set at the time a person is allowed for benefits, which is referred to as the comparison point decision (CPD) in the context of CDRs.

Each year, SSA runs predictive models to determine the likelihood that a beneficiary would be ceased for benefits if they underwent a medical CDR. For those beneficiaries whose diary date has come due, the scores from the predictive model are then used to determine next steps in the CDR process. Beneficiaries with High predictive model scores (probability above 4.21%) will undergo a full medical review (FMR) where beneficiaries will complete the Continuing Disability Review Report (Form SSA-454-BK) and recent medical evidence will be collected from providers to compare with CPD evidence under the Medical Improvement Review Standard (MIRS). If the beneficiary is found to have medical improvement related to the ability to work and no new impairments prevent the individual from performing SGA, the beneficiary will have benefits ceased. Beneficiaries with Low predictive model scores (probability below 2.01%) will be sent a Disability Update Report (Form SSA-455), also known as the CDR mailer. Based on the information received from the CDR mailer, a beneficiary will either continue on benefits or be sent for a FMR. Note that a beneficiary's benefits can only be ceased due to findings from a FMR. Beneficiaries with Medium predictive model scores (probability between 2.01% and 4.20%) are generally sent a CDR mailer although some cases are directly released for FMR depending on the resources SSA has available for CDRs within a given fiscal year.

The effectiveness of the CDR business process is partially driven by whether resources are available for processing CDR cases. Historically, SSA had faced backlogs in processing CDR cases. This backlog was cleared in FY 2018; however, no data have been published on the CDR backlog since [4]. The number of CDR cases processed dropped from 2,286,326 in FY 2018 to 1,506,195 in FY 2022 with the fewest CDR cases processed in FY 2020 during the COVID-19 pandemic (1,492,926 cases). There has also been a downward trend in initial cessation rates over time, dropping from 27.1% in FY 2013 to 21.5% in FY 2022. The FY 2025 President's Budget includes funding for 575,000 FMRs each in FYs 2024 and 2025 [5], which is 8,000 fewer than were conducted in FY2022 and 311,000 fewer FMRs than were conducted in FY 2018 when SSA caught up on the backlog of CDR cases. This is an improvement over FY 2023, though,

which saw 550,197 FMRs completed. With limited funding available for conducting FMRs, SSA will need to seek new approaches for maximizing the effectiveness of the CDR program.

Under the MIRS, an adjudicator must assess whether there has been improvement in the signs, symptoms, or laboratory findings of the CPD's medically determinable impairments. According to the guidance provided on evaluating medical improvement, "symptoms, signs, and laboratory findings include any evidence of physical or mental functioning used with medical findings at the time of the CPD" [6]. Having a consistent, comparable, standardized measure of physical and mental functioning that can be used throughout the CDR process – from case selection to FMR assessment – offers the potential for new efficiencies and advancements in SSA's business processes.

The Work Disability Functional Assessment Battery (WD-FAB)

A multi-year collaboration among the National Institutes of Health (NIH), Boston University, and SSA resulted in the development of the Work Disability Functional Assessment Battery (WD-FAB), an instrument to measure functional abilities that are related to work. The WD-FAB uses item response theory (IRT), a more modern test theory, that allows for a person's ability in a given area to be estimated and re-estimated with every question asked. Leveraging IRT methods, the WD-FAB also incorporates a computer adaptive testing (CAT) algorithm to select questions that will be most informative for estimating a person's ability and to introduce stopping rules, so a respondent does not have to keep answering questions once the desired reliability or maximum number of items is reached. These technologies allow for the WD-FAB to be an efficient and adaptable measure of a person's functional abilities.

The WD-FAB output consists of scores and standard errors for eight scales across the domains of physical and mental functioning. The scales corresponding to the physical functioning domain include Basic Mobility (BM), Upper Body Function (UBF), Fine Motor Function (FMF), and Community Mobility (CM). The scales for the mental functioning domain consist of Communication & Cognition (CC), Self-Regulation (SR), Resilience & Sociability (RS), and Mood & Emotions (ME). Each scale is normalized to a mean score of 50 and standard deviation (SD) of 10 with lower scores indicating lower functional ability and higher scores indicating higher functional ability. While the WD-FAB's CAT algorithm requires a respondent to answer specific questions to calculate each scale score, the question responses are not part of the WD-FAB output. This is similar to other measurement instruments that provide summary scores. To aid in score interpretation, the scale scores correspond to particular functional levels, which were established using a modified-Delphi process [7]. There are five functional levels for the physical functioning scales – Lowest, Low, Average, High, Highest – ranging from persistent and significant limitations to no limitations. For the mental functioning scales, there are four functional levels – Lowest, Low, Average, Above Average – again ranging from persistent and significant limitations to minimal or no limitation. Based on respondents' functional levels across all eight scales, we have defined functional profiles to summarize the number of functional limitations a person has. The five profiles are unique, cover all possible respondents, and are ranked ordinally from most functional limitations to fewest:

Profile 1 (P1): Multiple Significant Limitations (2 or more Lowest/Low scores)
Profile 2 (P2): One Significant Limitation (Exactly 1 Lowest/Low score)
Profile 3 (P3): Multiple Limitations (2 or more Average scores with no Lowest/Low)

Profile 4 (P4): Single Limitation (Exactly 1 Average score with no Lowest/Low)
Profile 5 (P5): Minimal to no Limitations (All scores greater than Average)

Thus, just like the WD-FAB scale scores, the higher the number of the functional profile, the greater the functional ability. The WD-FAB scale scores, functional levels, and functional profiles are all considered part of the WD-FAB data that can be used across research and practical applications.

The development of the WD-FAB followed a scientifically rigorous process consistent with an established framework to develop and test questions and establish the psychometric properties of the instrument [8]. Before the WD-FAB can be deployed in an applied setting, though, it must be tested in the environment of intended use to understand its role and impact. In 2018, SSA expressed interest in considering the use of the WD-FAB in the CDR process. The NIH initially designed a multi-year longitudinal study for SSA that would assess the correspondence between WD-FAB data and CDR diary designations. In response, SSA requested a smaller pilot study be designed that would instead focus on associations between CDR measures (i.e., the CDR mailer and predictive model scores) and WD-FAB data that could be used as a foundation for future studies to characterize the utility of the WD-FAB in the CDR process more fully. This led to the development of the WD-FAB Research Study.

The WD-FAB Research Study

The WD-FAB Research Study was a pilot study designed to examine the feasibility and value of integrating the WD-FAB into SSA's CDR business process and to look at associations between the current CDR business process and WD-FAB data from a single time point as well as longitudinally to assess change in WD-FAB data. Due to study restrictions specified by SSA, WD-FAB data would need to be collected six months apart, even though a six-month time frame does not correspond to any current CDR business process. The NIH provided SSA with a study design for this pilot study at the end of FY 2019 [9]. SSA then contracted with Westat, a survey research firm, for the survey data collection, which began in July 2022 and ran until July 2023.

An initial participant pool of 6,460 beneficiaries with CDR diaries due in FY 2022 were identified using a stratified sampling approach that took into consideration diary type, CDR predictive model category, and age. However, after Westat faced challenges in recruiting the target sample, an additional 10,000 beneficiaries who already had their cases released for full medical review were added to the participant pool. From this full pool of 16,460 beneficiaries, 2,406 (15%) beneficiaries completed the first survey and WD-FAB administration. After six months, these beneficiaries were recontacted for the second survey data collection and WD-FAB administration. Of the 2,406 beneficiaries who participated in the first survey, 1,603 (67%) beneficiaries completed the second WD-FAB administration, as well.

In parallel to the Westat data collection efforts, SSA provided the NIH with CDR administrative data and medical evidence for beneficiaries from the sample with an initial FMR decision. Of the beneficiaries who completed both WD-FAB administrations, 1,263 (79%) had a CDR decision of either continuance or cessation with an initial cessation rate of 8%. For those cases with a cessation, 19 (19%) were ceased due to failure to cooperate and were excluded from analysis. Thus, our final analytic sample consisted of 1,244 beneficiaries. Table 1 summarizes the characteristics of this sample using SSA administrative data.

Table 1. SSA Administrative Data for Analytic Sample

SSA Administrative Data	Mean / Count	SD / Percent
Age at CDR (Years)	47.3	11.5
Time on Rolls (Years)	4.0	1.9
Program Type		
Title II	748	60.1
Title XVI	450	36.2
Concurrent	46	3.7
Diary Type		
MIE	348	28.0
MIP	738	59.3
MINE	158	12.7
Predictive Model Score		
Low	260	20.9
Medium	280	22.5
High	704	56.6
CPD Impairment Type		
Mental	459	36.9
Physical	785	63.1
CDR Impairment Type		
Mental	440	35.4
Physical	804	64.6
Initial CDR Decision		
Continuance	1162	93.4
Cessation	82	6.6

Compared to non-participants, i.e., beneficiaries who did not respond or were found to be ineligible to take the survey, the analytic sample was older (48.0 years old vs. 44.9 years old) and more likely to be female (57.6% vs. 46.5%), on SSI benefits only (60.1% vs. 46.0%), allowed at CPD for a physical impairment (63.1% vs. 56.0%), and less likely to have a diary of Medical Improvement Expected (28.0% vs. 31.2%) or a high predictive model score (56.6% vs. 64.2%). See Table A1 in Appendix A for more details.

As part of the survey administration, Westat collected a number of demographic variables including sex, race and ethnicity, marital status, and highest level of education. The mode of administration – either telephone or web – was also recorded for each respondent. Figure 1 summarizes the demographic characteristics of the analytic sample. For each demographic category, the figure shows the proportion for the total analytic sample in blue, as well as by administration mode with proportion of telephone respondents in orange and web respondents in green. Comparing across administration mode, those who participated via phone were more likely to be male (44.0% vs. 42.2%), Black or African American (33.3% vs. 17.8%), divorced (28.5% vs. 17.3%), have at most a high school education (50.1% vs. 46.0%), and to have a physical primary impairment at CPD and CDR (71.2% vs. 60.4% and 73.8% vs. 61.6%, respectively) compared with those who completed the surveys over the internet. More details are included in Table A2 in Appendix A.

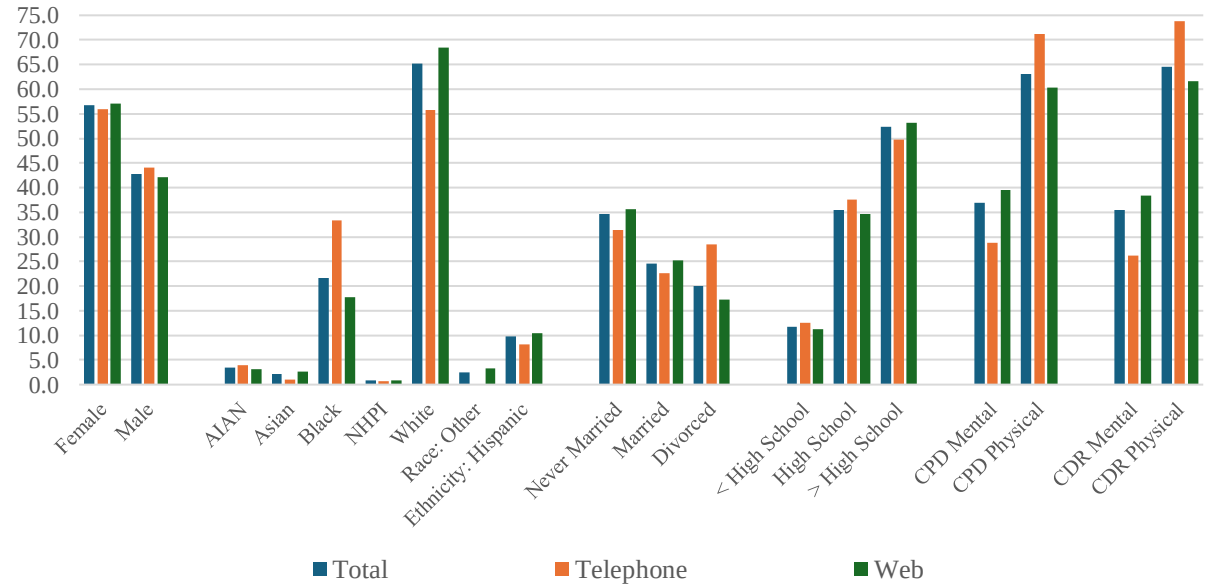

Figure 1. Analytic Sample Demographics[1] *by Administration Mode*

This report includes analyses of the functioning information available in the medical evidence as extracted using Natural Language Processing (NLP) techniques, and how it relates to WD-FAB data and CDR business process artifacts. From the 1,244 beneficiaries in the final analytic sample, we selected a subset of 452 for whom we had access to the medical evidence at both the CPD and CDR timepoints. In this NLP analytic sample, 44 beneficiaries (10%) had a CDR decision of cessation. The medical evidence for these cases consisted of 2,579 documents for the CPD timepoint, 1,868 for the CDR timepoint, and 441 Disability Determination Explanation (DDE) documents at CDR, or an average of 5.71 CPD, 4.13 CDR, and 0.98 DDE documents per case. The average length of available evidence was 36,876 sentences at CPD (SD = 50,283), 29,947 sentences at CDR (SD = 114,522), and 427 sentences in DDE documents (SD = 302). In the NLP analyses described in this report, DDE documents are excluded from analysis except when explicitly stated otherwise.

In the following sections, we will describe the analytic methods and findings to address each of the WD-FAB Research Study objectives along with summary recommendations based on these findings.

Utility of the WD-FAB in the CDR Process

The original study design proposed three objectives for exploring the potential application, utility, and impact of the WD-FAB in the CDR process:

1. Assess single WD-FAB administrations within the CDR process.
2. Assess change in WD-FAB data beyond minimal detectable change.
3. Examine the feasibility and value of integrating the WD-FAB into the CDR business process.

[1] Abbreviations used in race category: American Indian or Alaska Native (AIAN); Black or African American (Black); Native Hawaiian or Other Pacific Islander (NHPI); All other responses (Race: Other)

These objectives included characterizing beneficiaries using WD-FAB data and comparisons with CDR data and outcomes in order to understand associations between the WD-FAB and current CDR processes, as well as to identify areas where the WD-FAB can contribute as an additional data source. The methods and findings for these study objectives are presented below.

Single WD-FAB Administration

As described above in the introduction to the WD-FAB, WD-FAB data consist of the individual scale scores and standard errors, along with the corresponding functional levels, which are then used to define the functional profiles (see page 9). When looking at single administration data to characterize beneficiaries and compare WD-FAB data to CDR measures, we included both the scale scores and the functional profiles as part of the WD-FAB data in our analyses.

Characterizing Beneficiaries Using Single Administration WD-FAB Data

Single administration WD-FAB data consists of data from the WD-FAB administrations at each time point – Survey 1 and Survey 2. To characterize beneficiaries using these data, we looked at the distribution of the WD-FAB scores and functional profiles at each time point, for the total analytic sample and by sub-population defined by SSA administrative data such as program and impairment type.

The distributions of WD-FAB scores were similar between Survey 1 and Survey 2 for both the total analytic sample and for each sub-population. Figure 2 displays box plots with mean, interquartile range, minimum, maximum, and outlier scores for each WD-FAB scale at the two survey administrations. Mean WD-FAB scores for the analytic sample tended to be between 45 and 51, indicating at least some limitation in these areas. The scores for Fine Motor Function tended to be higher with mean scores closer to 62. This is consistent with previous studies of the WD-FAB [10].

Figure 2. Box Plot of WD-FAB Scores at Survey 1 (S1) and Survey 2 (S2)

Comparing across sub-populations, those on SSI or SSDI only tend to have similar mean scores across the WD-FAB scales, while those who are on both SSI and SSDI have higher physical functioning scores on average. As we would expect, individuals with a physical primary impairment, either at CPD or CDR, have lower scores in scales such as Basic Mobility and Upper Body Function, while those with a mental primary impairment have lower scores in the four mental functioning scales with the largest difference being in Mood & Emotions. When we look at the three most common body systems corresponding to a physical primary impairment at the time of the CDR – Body System (BS) 1: Musculoskeletal, BS 11: Neurological, and BS 13: Cancer – beneficiaries with a Musculoskeletal disorder had lower Basic Mobility and Upper Body Function scores on average, while those with a Neurological disorder tended to have lower Fine Motor Function scores. See Tables B1 and B2 in Appendix B for further details.

Similarly, the distributions of WD-FAB functional profiles were consistent across the two survey administrations for the total analytic sample (Figure 3). While there are five profile categories, the majority of the sample had functional profiles corresponding to the first three categories, which are indicative of the most functional limitations. In general, beneficiaries did not have profiles in the two highest functioning categories – Profile 4 and Profile 5 – other than two beneficiaries (0.2%) with Profile 4 at Survey 1.

Figure 3. Distribution of WD-FAB Profiles at Survey 1 and Survey 2

The distributions across sub-populations showed greater variability (Figure 4 and Figure 5). SSI beneficiaries tended to fall more into Profile 1 or Profile 3, whereas SSDI and Concurrent beneficiaries had a higher proportion of Profile 1 corresponding to the lowest functional abilities. The majority of beneficiaries with a mental primary impairment fell into Profile 1, while those with a physical primary impairment tended to have a higher proportion of Profile 3. We saw different distributions across the three most common physical body systems; there were more beneficiaries in Profiles 1 and 3 for Musculoskeletal (CDR BS1) and Neurological (CDR BS11) disorders while the majority of beneficiaries with a Cancer (CDR BS13) primary impairment fell into Profile 3 indicating the presence of multiple limitations, although not the most significant limitations. These distributions are also provided in Table B3 in Appendix B.

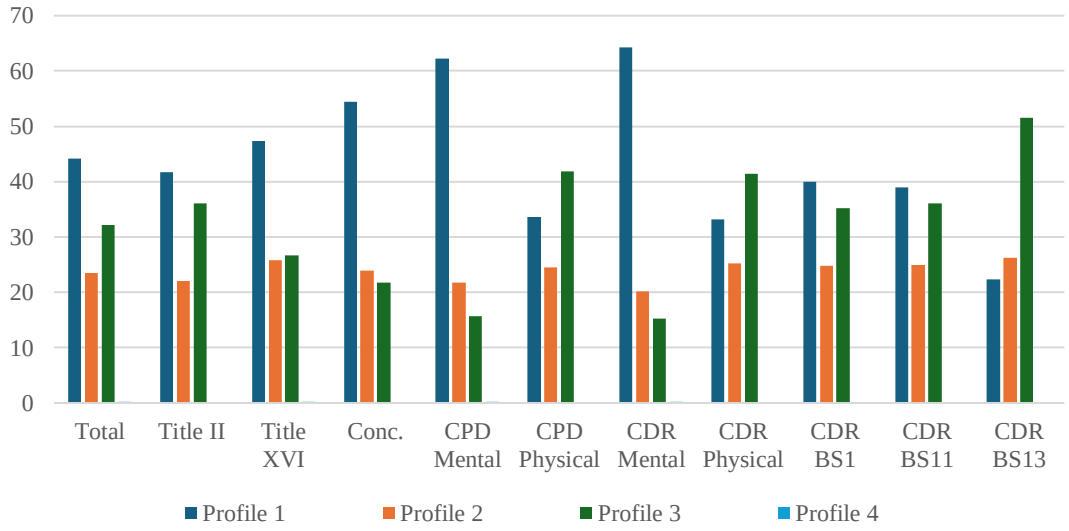

Figure 4. Distribution of WD-FAB Profiles by SSA Sub-Population (Survey 1)

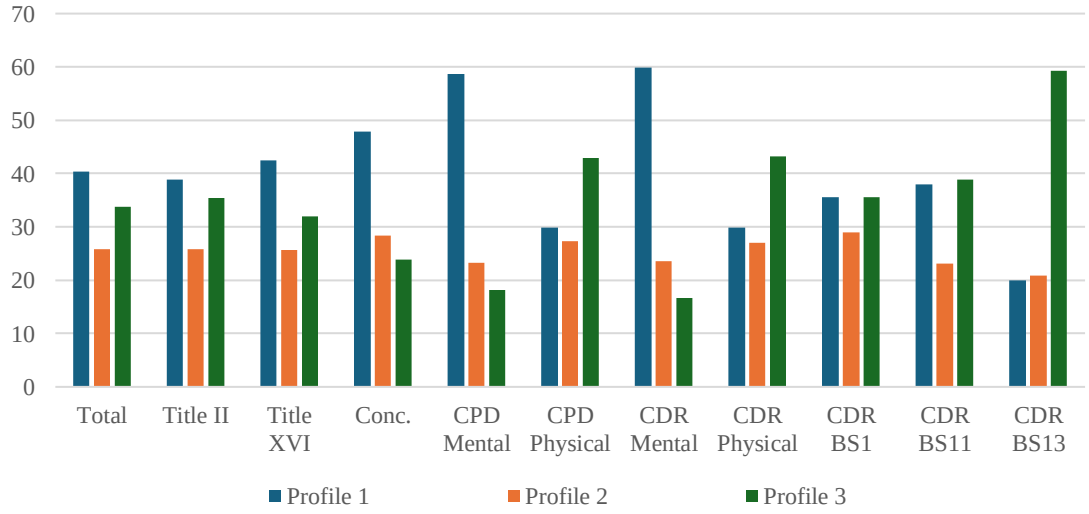

Figure 5. Distribution of WD-FAB Profiles by SSA Sub-Population (Survey 2)

Comparing Single Administration WD-FAB Data to CDR Measures

When comparing WD-FAB data to CDR measures, we include data from across the various aspects of the CDR business process: CDR diary dates, predictive model scores, Form SSA-455 (CDR mailer) questions, and information from the FMR such as Form SSA-454 Daily Activities and medical evidence.

The CDR diary dates consist of three types – Medical Improvement Expected (MIE), Medical Improvement Possible (MIP), and Medical Improvement Not Expected (MINE). Table 2 shows the mean WD-FAB scores for these categories at both Survey 1 and Survey 2. Scores are consistent between the two survey administrations with those with a MINE diary designation generally having higher scores on average.

Table 2. Mean WD-FAB scores by CDR Diary Date Category

Diary Date	BM[2]	UBF	FMF	CM	CC	SR	RS	ME
Survey 1								
MIE	48.3	45.5	62.1	44.3	46.8	50.3	44.0	48.7
MIP	48.7	46.2	62.7	45.8	47.7	51.4	45.4	49.9
MINE	49.2	46.7	62.1	47.5	48.9	53.0	48.1	53.9
Survey 2								
MIE	48.6	45.6	62.2	45.2	47.5	50.7	44.2	49.6
MIP	48.8	46.4	62.5	45.8	47.9	51.9	45.7	50.7
MINE	48.9	46.9	62.5	47.5	49.0	53.1	47.9	54.8

Comparing using the WD-FAB functional profiles, we again see consistency in the distribution across the two survey administrations (Figure 6). About half of the beneficiaries with a MIE diary fall into the first profile category, Profile 1, consisting of individuals with multiple significant limitations. Nearly half of those with a MINE diary fall into the third profile category, Profile 3 corresponding to multiple limitations.

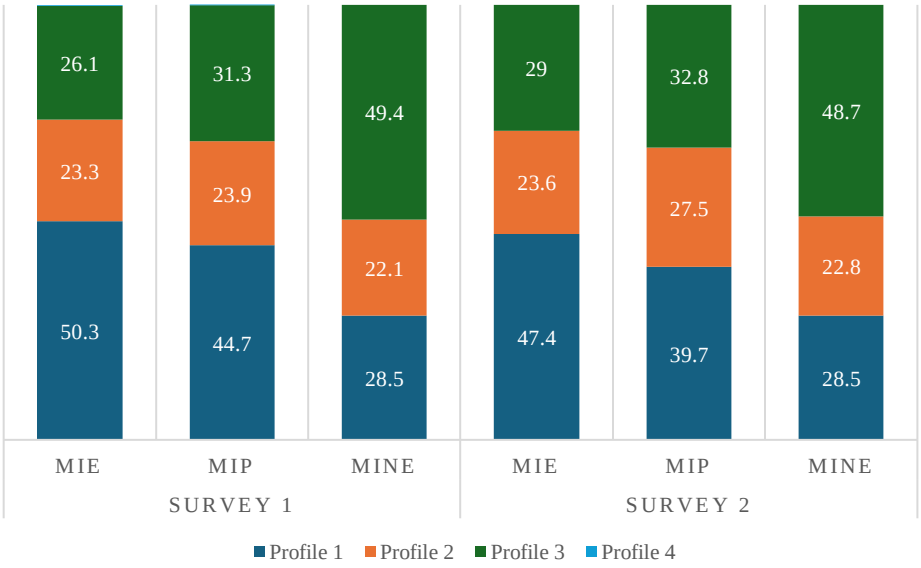

Figure 6. Distribution of WD-FAB Profiles by CDR Diary Type (Survey 1 and 2)

CDR predictive model scores are stratified into three categories – Low, Medium, High – based on likelihood of cessation.
Table 3 shows the mean WD-FAB scores for these three categories at both Survey 1 and Survey 2. For the physical functioning scales, we see increasing mean scores as the predictive model scores increase from Low to High, suggesting that higher physical functional abilities are associated with higher predictive model scores and thus higher likelihood of cessation. There is no consistent trend or meaningful difference in mean scores for the mental functioning scales across predictive model categories.

[2] BM: Basic Mobility; UBF: Upper Body Function; FMF: Fine Motor Function; CM: Community Mobility; CC: Communication & Cognition; SR: Self-Regulation; RS: Resilience & Sociability; ME: Mood & Emotions

Table 3. Mean WD-FAB scores by Predictive Model Score Category

CDR PMS	BM[3]	UBF	FMF	CM	CC	SR	RS	ME
Survey 1								
Low	46.1	44.2	60.8	44.4	47.3	51.7	45.9	49.8
Medium	46.8	44.3	61.3	44.3	47.0	50.3	44.6	49.4
High	50.3	47.5	63.5	46.5	48.0	51.6	45.4	50.4
Survey 2								
Low	46.3	44.6	61.3	44.3	47.8	52.2	46.4	51.4
Medium	47.1	44.9	61.6	45.1	47.3	51.1	44.8	49.9
High	50.3	47.4	63.2	46.7	48.2	51.9	45.6	51.1

Figure 7 shows the distribution of the WD-FAB functional profiles by CDR predictive model score (PMS) category for Survey 1 and Survey 2. We see a slight decrease in the proportion of beneficiaries in Profile 1 between the two survey administrations across all three categories indicating a shift towards slightly greater functional abilities, although Profile 1 still has the highest proportion for all three predictive model score categories.

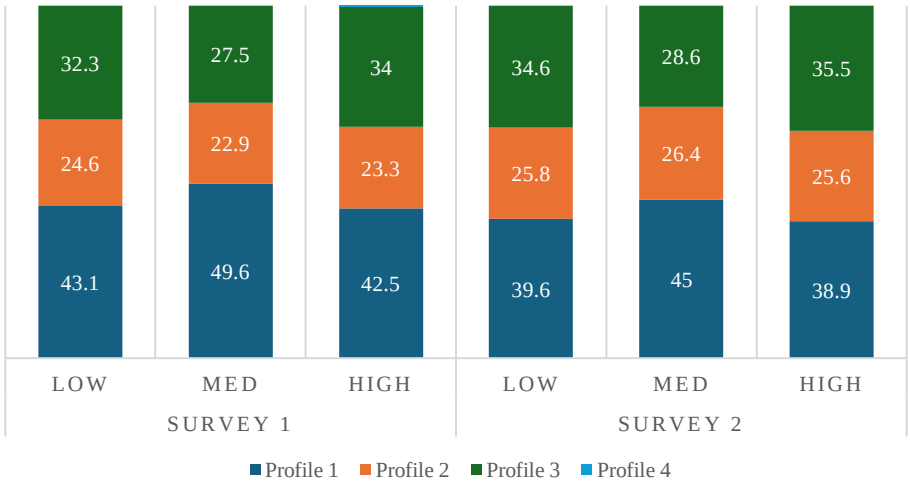

Figure 7. Distribution of WD-FAB Profiles by Predictive Model Score (Surveys 1 and 2)

We used logistic regression to model the relationship between the WD-FAB scores and CDR predictive model scores. For this analysis, we coded the High, Medium, and Low model score categories as 3, 2, and 1, respectively. In particular, we looked at the proportional odds to define the relationship: Predictive Model Score ~ (BM + UBF + FMF + CM + CC + SR + RS + ME). For both Survey 1 and Survey 2 data, Basic Mobility scores were found to have a significant relationship with the predictive model scores. For Survey 1, Basic Mobility had an odds ratio of 1.0593 (95% confidence interval (1.0252, 1.0953)), meaning the chance of a higher predictive model score is 1.0593 times higher as the Basic Mobility score increases by 1 point. Similarly for Survey 2, Basic Mobility has an odds ratio of 1.0513 (95% confidence interval (1.0181, 1.0863)).

[3] BM: Basic Mobility; UBF: Upper Body Function; FMF: Fine Motor Function; CM: Community Mobility; CC: Communication & Cognition; SR: Self-Regulation; RS: Resilience & Sociability; ME: Mood & Emotions

To ensure that data would be available from both the CDR mailer and FMRs for study participants, the CDR mailer questions were collected by Westat as part of both survey administrations. The mailer questions and respective response options are included in Table 4. Note that the mailer questions were mislabeled in the Westat codebook and data as SSA445 rather than the correct form number (SSA455). We use the Westat labels in this report for consistency with the provided data.

Table 4. CDR Mailer (SSA-455) Questions and Response Options

Variable Name	Question Statement	Response Options
SSA445_1	Within the last 2 years have you worked for someone or been self-employed?	Yes; No
SSA445_2	Describe your health status within the last 2 years.	Better; Same; Worse
SSA445_3	Within the last 2 years has your doctor told you that you can return to work?	Yes; No
SSA445_4	Within the last 2 years have you attended any school or work training program(s)?	Yes; No
SSA445_5	Would you be interested in receiving rehabilitation or other services that could help you get back to work?	Yes; No
SSA445_6	Within the last 2 years have you been hospitalized or had any surgery?	Yes; No
SSA445_7	Within the last 2 years have you gone to a doctor or clinic for your condition?	Yes; No

Given the multiple categorical questions associated with the CDR mailer, we used logistic regression models to look at the relationship between the WD-FAB scale scores and mailer responses. For this modeling, each question was considered to be a dependent variable. The predictor variables included WD-FAB scores at each time point. Regression models characterize the relationship between the predictor variables and the dependent variable. The results for Survey 1 are shown in Table 5. Higher Upper Body Function and Resilience & Sociability scores were associated with a higher chance of having worked in the last 2 years (SSA445_1). Higher scores in Basic Mobility, Upper Body Function, and Mood & Emotions, or lower scores in Self-Regulation led to a higher chance of having same or better health (SSA445_2). Higher Basic Mobility and Self-Regulation scores led to a higher chance of a doctor recommending the beneficiary return to work (SSA445_3). There were no significant associations with the other CDR mailer questions for the WD-FAB data from Survey 1.

Table 5. Logistic Regression Comparing WD-FAB to CDR Mailer (Survey 1)

Survey 1

SSA445_1: Worked in Last 2 Years				
	Estimate	Std. Error	t-value	Pr(>|t|)
Upper Body Function	0.0084	0.0041	2.0700	0.0389
Resilience & Sociability	0.0091	0.0035	2.6380	0.0086
SSA445_2: Health Last 2 Years				
	Estimate	Std. Error	t-value	Pr(>|t|)

Basic Mobility	0.0094	0.0037	2.5400	0.0113
Upper Body Function	0.0129	0.0044	2.9570	0.0032
Self-Regulation	-0.0095	0.0036	-2.6670	0.0079
Mood & Emotions	0.0100	0.0028	3.5550	0.0004
SSA445_3: Doctor Said Return to Work Last 2 Years				
	Estimate	Std. Error	t-value	Pr(>|t|)
Basic Mobility	0.0078	0.0026	2.9560	0.0032
Resilience & Sociability	0.0059	0.0027	2.2420	0.0253

The results for Survey 2 are shown in Table 6. Again, higher Upper Body Function and Resilience & Sociability scores were associated with a higher chance of having worked in the last 2 years (SSA445_1). Similarly, higher scores in Basic Mobility, Upper Body Function, and Mood & Emotions led to a higher chance of having same or better health (SSA445_2). However, for Survey 2, lower scores in Communication & Cognition rather than Self-Regulation led to a higher chance of having same or better health. Higher Upper Body Function scores led to a higher chance of a doctor recommending the beneficiary return to work at Survey 2 (SSA445_3). Another difference from Survey 2 was that higher scores in Basic Mobility and Resilience & Sociability scales were associated with a lower chance of having visited a doctor or clinic in the past 2 years. Tables B4 and B5 in Appendix B include the mean WD-FAB scale scores and distribution of the functional profiles for the response options to the four CDR mailer questions that showed an association with the WD-FAB data to underscore the findings from the logistic regression models.

Table 6. Logistic Regression Comparing WD-FAB to CDR Mailer (Survey 2)

Survey 2

SSA445_1: Worked in Last 2 Years				
	Estimate	Std. Error	t-value	Pr(>|t|)
Upper Body Function	0.0142	0.0040	3.5380	0.0004
Resilience & Sociability	0.0079	0.0034	2.3450	0.0194
SSA445_2: Health Last 2 Years				
	Estimate	Std. Error	t-value	Pr(>|t|)
Basic Mobility	0.0115	0.0037	3.1520	0.0017
Upper Body Function	0.0137	0.0043	3.2120	0.0014
Communication & Cognition	-0.0122	0.0050	-2.4400	0.0150
Mood & Emotions	0.0127	0.0028	4.4960	0.0000
SSA445_3: Doctor Said Return to Work Last 2 Years				
	Estimate	Std. Error	t-value	Pr(>|t|)
Upper Body Function	0.0094	0.0031	3.0810	0.0022
SSA445_7: Doctor/Clinic for Conditions Last 2 Years				
	Estimate	Std. Error	t-value	Pr(>|t|)

Basic Mobility	-0.0079	0.0019	-4.2160	0.0000
Resilience & Sociability	-0.0037	0.0018	-2.0130	0.0445

For the FMR, SSA collects data from beneficiaries via Form SSA-454-BK, the Continuing Disability Review Report, and medical records from providers to compare with evidence from the CPD. For these analyses, we compared WD-FAB data to self-reported daily activity limitations from Form SSA-454, as well as mentions of function from the medical evidence that we extracted using our natural language processing (NLP) models for identifying functioning information. In order to compare self-reported daily activity limitations to WD-FAB data, we mapped each activity to the corresponding WD-FAB scale. We then calculated the distribution of WD-FAB scale scores across the mean reported difficulty levels for the daily activities that corresponded to that scale. For example, there were four daily activities that mapped to the Basic Mobility scale – Walking, Standing, Sitting, and Doing chores – which correspond to five possible mean reported difficulty values – no difficulty in any of the four activities (0.0), difficulty in one activity (0.25), difficulty in two activities (0.5), difficulty in three activities (0.75), and difficulty in all four daily activities (1.0). As the number of reported difficulties increased, the lower the Basic Mobility scores were on average (Figure 8). We observed the same trend across all WD-FAB scales. Box plots of the distributions for the remaining seven scales are included in Figures B1-B7 Appendix B.

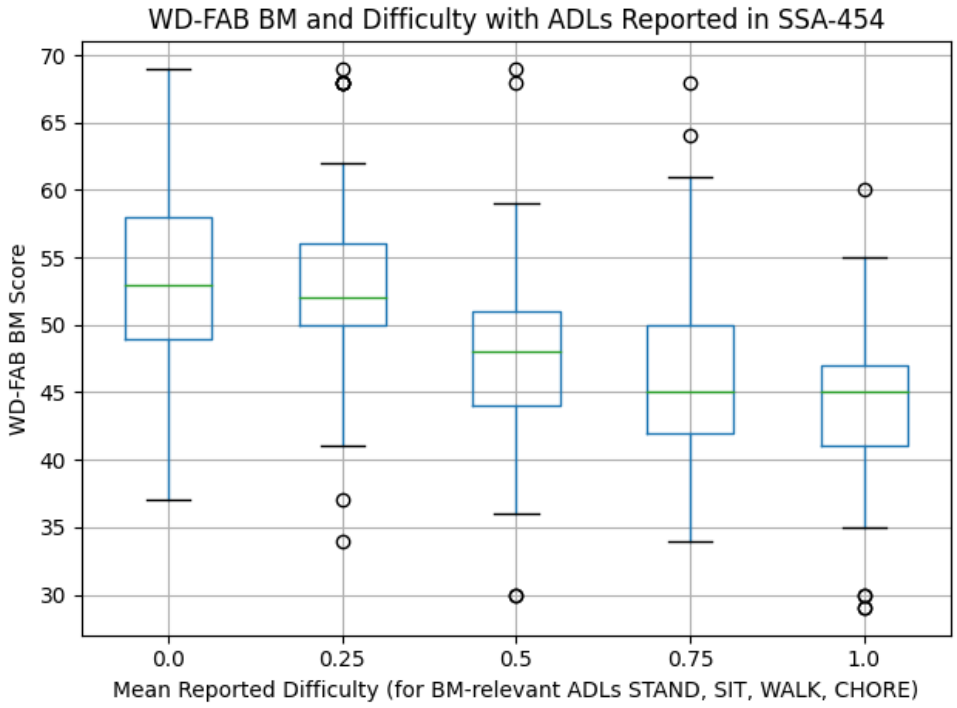

Figure 8. Box Plot of WD-FAB Basic Mobility Scores by Mean Reported Difficulty in SSA-454 Daily Activities

In order to compare WD-FAB data to medical evidence, we ran our NLP models for identifying functioning information on the records available from the CPD and CDR cases for a subset of the

analytic sample. When analyzing single administration WD-FAB data, we focused on the CDR medical evidence as more relevant to the administration time point. First, we looked at the number of CDR cases with relevant functioning information stratified by WD-FAB scale functional level to establish how much evidentiary support would be available for analysis at each level (Figure 9). Cases with physical functioning information tended to be concentrated in the High and Highest functional levels, although we did identify cases across all five functional levels. Cases with Low to Average mental functioning scores were found to have relevant functioning information available in the medical record.

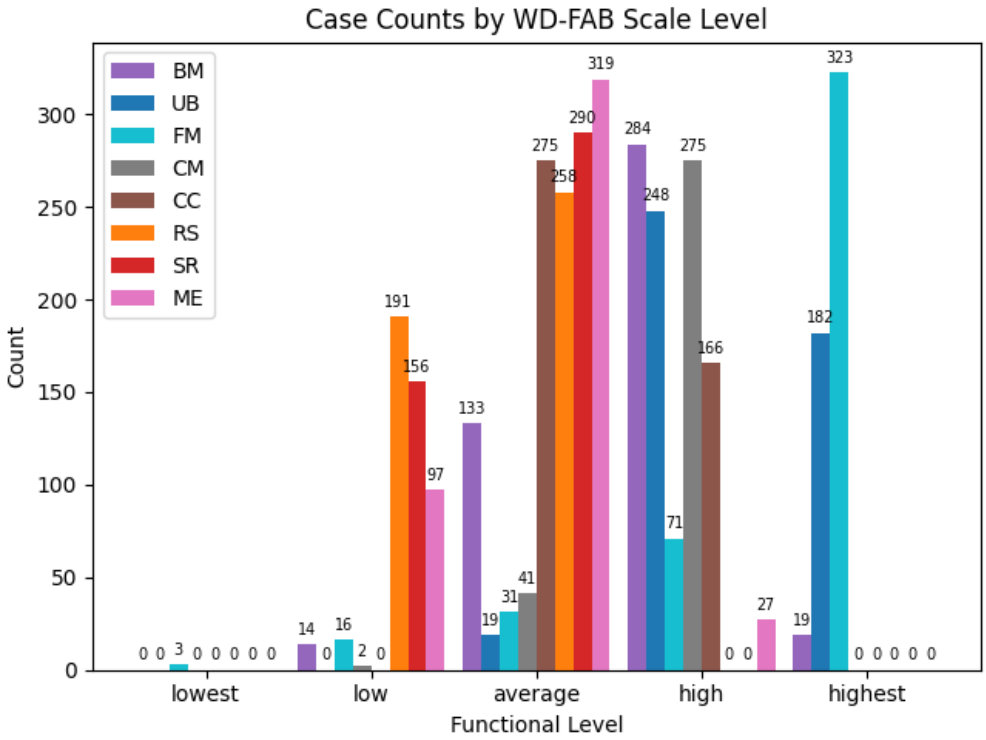

Figure 9. Number of CDR cases with Functioning Information Across WD-FAB scales by WD-FAB Functional Level

We then normalized the case counts to case presence, i.e., the proportion of cases within a given WD-FAB functional level that have relevant functioning information present in the medical evidence (Figure 10). We excluded the Lowest functioning level because there were not enough cases to support analysis. For the majority of the WD-FAB scales, the case presence decreased as functional level increased, which is the expected relationship with higher functional ability less likely to be documented in the medical record. Six of the eight WD-FAB scales had high case presence ratios (greater than 0.8) across all functional levels. There were a smaller proportion of cases with functioning information for Community Mobility and Fine Motor Function when the functional level was Average or higher, although relevant functioning information was still identified in over half the cases.

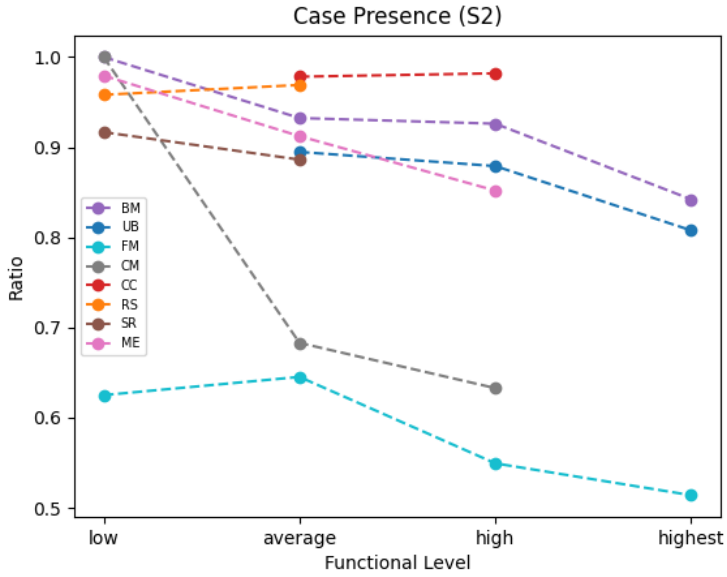

Figure 10. Proportion of Cases with Functioning Information by WD-FAB Functional Level

The amount of functioning information available varied more widely by WD-FAB scale (Figure 11). The Communication & Cognition scale tended to have the most information available at 354.5 relevant mentions per case. We excluded this scale from the figure to improve interpretability of results for the other seven scales. The scales of Basic Mobility, Resilience & Sociability, and Mood & Emotions tended to have over 30 mentions of relevant functioning information for cases with High functional abilities or below. The other three physical functioning scales of Upper Body Function, Fine Motor Function, and Community Mobility tended to have 1-10 mentions on average across all functional levels.

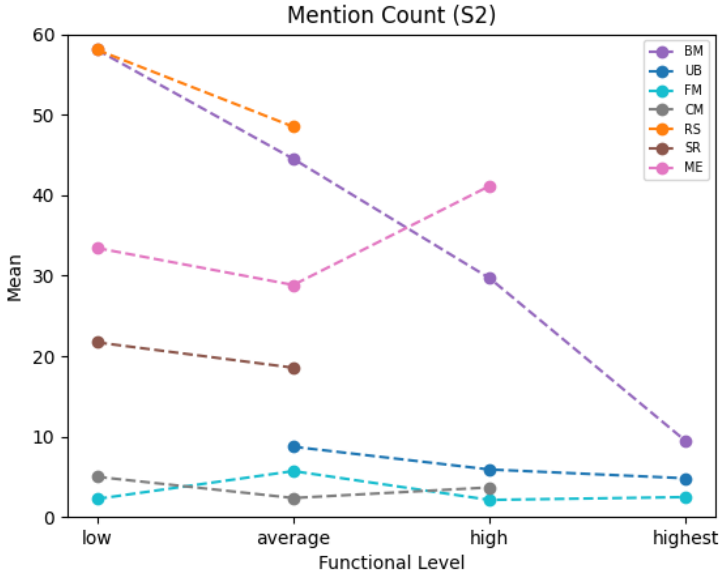

Figure 11. Mean Mention Count by WD-FAB Functional Level

Relating Single Administration WD-FAB Data to CDR Outcomes

In order to discuss associations between the WD-FAB data and CDR FMR outcomes, we first need to contextualize when the WD-FAB administrations took place in relation to the FMRs. Table 7 shows the mean time in days from the date of each survey administration to the date of the FMR decision. If the WD-FAB were to be used in the CDR process, then ostensibly the WD-FAB would be administered prior to the determination. However, we see that on average the study participants responded to Survey 1 four days after the FMR decision was already made. For those beneficiaries with a continuance, this decision was made even earlier, nearly nine days before responding to the first WD-FAB administration. When looking at cessation cases, though, we see that on average these beneficiaries completed Survey 1 two months before the FMR decision date. While we present associations between the WD-FAB data and FMR outcomes in this section, it should be noted that we are not comparing data from consistent time points, and thus, any associations we observe might be more related to the timing rather than the value of the outcome.

Table 7. Mean Time from WD-FAB Survey Administration to Decision Date in Days

	N	MEAN DAYS FROM SURVEY 1 TO DECISION (DCN – SURVEY 1)	MEAN DAYS FROM SURVEY 2 TO DECISION (DCN – SURVEY 2)
TOTAL	1244	-4.0	-194.5
CONTINUANCE	1162	-8.8	-199.2
CESSATION	82	63.6	-128.8

Looking at the distribution of WD-FAB scale scores for continuances (CO) vs. cessations (CE), we see that beneficiaries who were continued tend to have lower physical functioning scores, i.e., less functional ability, on average (Figure 12 and Figure 13). The distributions for the mental functioning for continuances vs. cessations are more similar. The mean WD-FAB scores from Survey 1 and Survey 2 for continuances vs. cessations are provided in Table B6 in Appendix B.

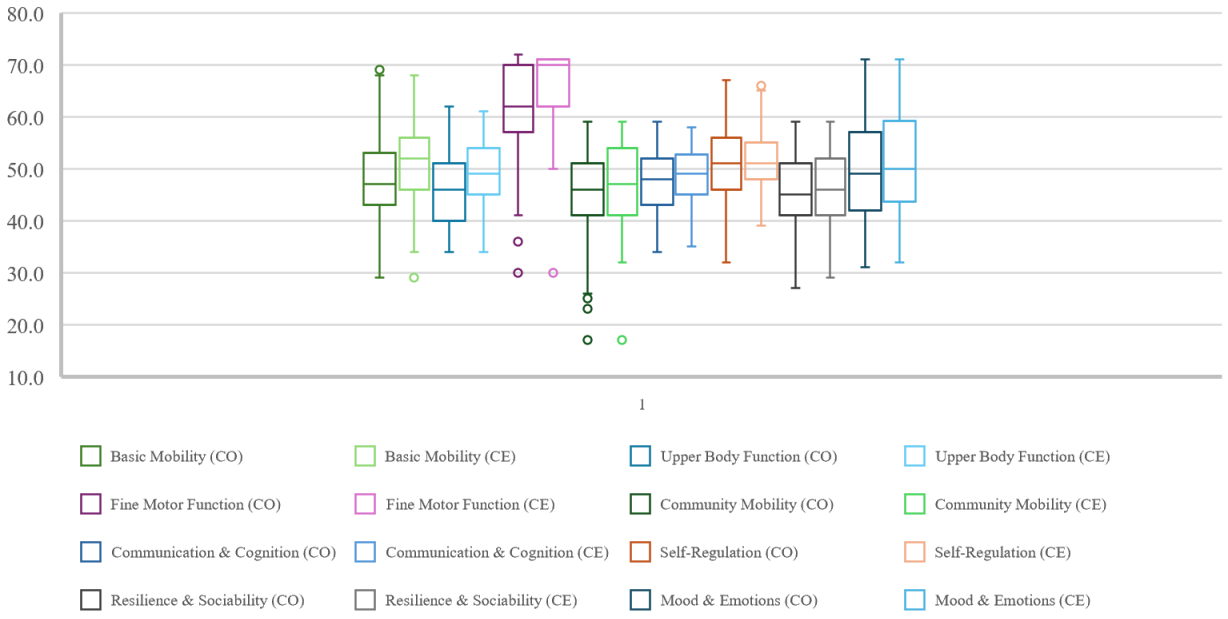

Figure 12. Box Plot of WD-FAB Scores for Continuances (CO) vs. Cessations (CE) at Survey 1

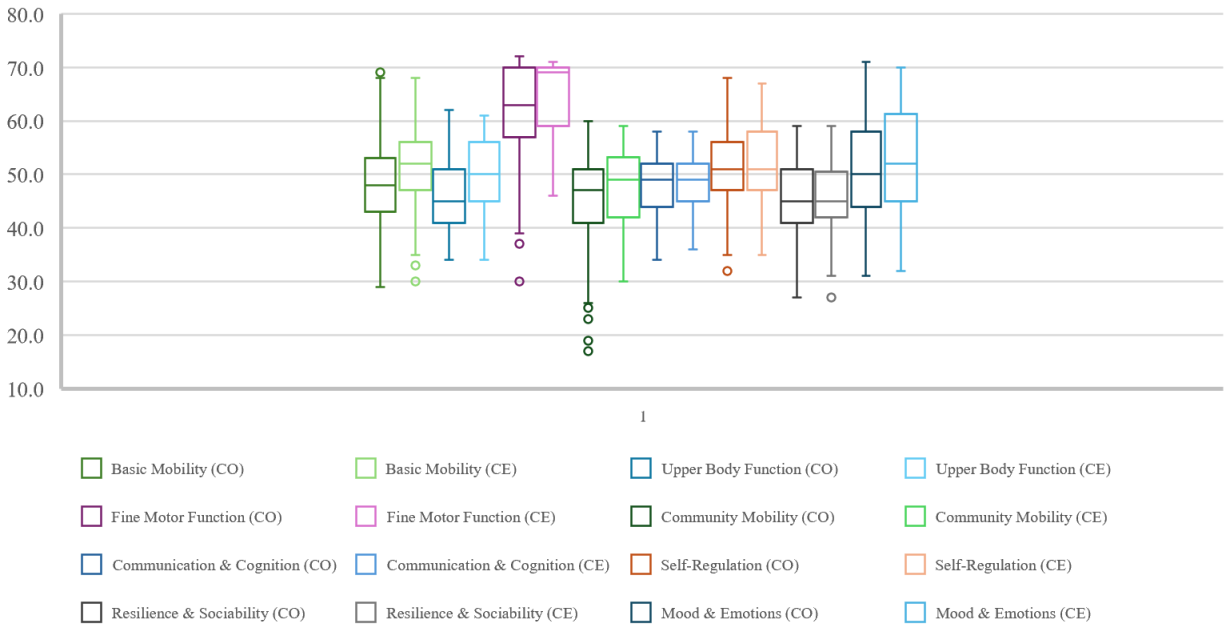

Figure 13. Box Plot of WD-FAB Scores for Continuances (CO) vs. Cessations (CE) at Survey 2

For the WD-FAB functional profiles, we looked at the intersection of the CDR predictive model scores and FMR outcomes since the predictive model scores are themselves an indicator of likelihood of cessation. Table 8 shows the distribution of the WD-FAB functional profiles across the predictive model score categories for both Survey 1 and Survey 2 along with the corresponding initial allowance rate (% CO), as well as total allowance rate for each WD-FAB profile.

Table 8. Distribution and Continuance Rate for WD-FAB Profiles by Predictive Model Score

	HIGH			**MEDIUM**			**LOW**			**TOTAL**
Survey 1	Count	Percent	% CO	Count	Percent	% CO	Count	Precent	% CO	% CO
Profile 1	299	42.5	92.6	139	49.6	97.8	112	43.1	98.2	95.1
Profile 2	164	23.3	86.0	64	22.9	96.9	64	24.6	96.9	90.8
Profile 3	239	33.9	89.5	77	27.5	97.4	84	32.3	98.8	93.0
Profile 4	2	0.3	100.0	0	0.0	N/A	0	0.0	N/A	100.0
Survey 2	Count	Percent	% CO	Count	Percent	% CO	Count	Precent	% CO	% CO
Profile 1	274	38.9	88.7	126	45.0	98.4	103	39.6	98.1	93.0
Profile 2	180	25.6	92.8	74	26.4	96.0	67	25.8	97.0	94.4
Profile 3	250	35.5	89.6	80	28.6	97.5	90	34.6	98.9	93.1

We also used logistic regression to analyze single point WD-FAB data in comparison to CDR outcomes. Cessations were coded as 1 and continuances were coded as 0. The predictor variables were the WD-FAB scores. The analyses were carried out for both Survey 1 and Survey 2 data.

For Survey 1, we did not find any significant relationships. For Survey 2, beneficiaries with higher Upper Body Function were more likely to have a cessation (Table 9).

Table 9. Logistic Regression Comparing WD-FAB to CDR Decision

	Full Medical Review Decision			
	Estimate	Std. Error	t-value	Pr(>|t|)
Survey 2 Upper Body Function	0.0989	0.0400	2.4710	0.0135

Change in WD-FAB Data

Given the central role that improvement plays in the CDR process, the main goal of this work was to look at functional change in the beneficiary population and how that change compares with CDR data. With the imposed study parameters, this pilot study only looks at functional change over a six-month period, which limits the associations that can be made with the CDR process. As part of the original study design, the aim was to have the WD-FAB data collections to occur approximately concurrently with the full medical reviews, however due to delays in data collection efforts, many beneficiaries in the sample ended up having the FMR decisions made prior to either WD-FAB administration, i.e., the period of functional change that the WD-FAB data measured occurred after the CDR case was closed and so does not align with actual CDR business processes or evidence used to make the CDR decision.

In the following sections, we will define change in WD-FAB data, characterize beneficiaries based on those changes, look at associations between change in WD-FAB data and CDR measures, and finally compare change to the FMR outcomes.

Characterizing Beneficiaries Based on Change in WD-FAB Data

Item response theory, the test methodology underlying the WD-FAB, provides a natural method to contextualize score changes relative to measurement precision. Every time a score estimate is reported, the WD-FAB also provides an estimated standard error on that score. The standard error of the change (SEC) in a scale (i.e., difference between two scores) is the square root of the sum of the squares of the respective score standard errors (SE):

$$SEC = \sqrt{SE_1^2 + SE_2^2}$$

Dividing score changes by their standard error yields a standardized difference. Figure 14 presents a scatterplot of the changes across all study participants, which looks at correlations in score changes between any two give WD-FAB scales. That is, whether change in function in one scale is correlated with change in function in another scale. Largely, only weak correlations between changes in the scale scores were present. Basic Mobility and Upper Body Function score changes were the most correlated (0.39), although still considered to be a weak correlation, and there were no other correlations exceeding 0.3. This indicates that improvement in any given scale is not predictive of improvement in other scales.

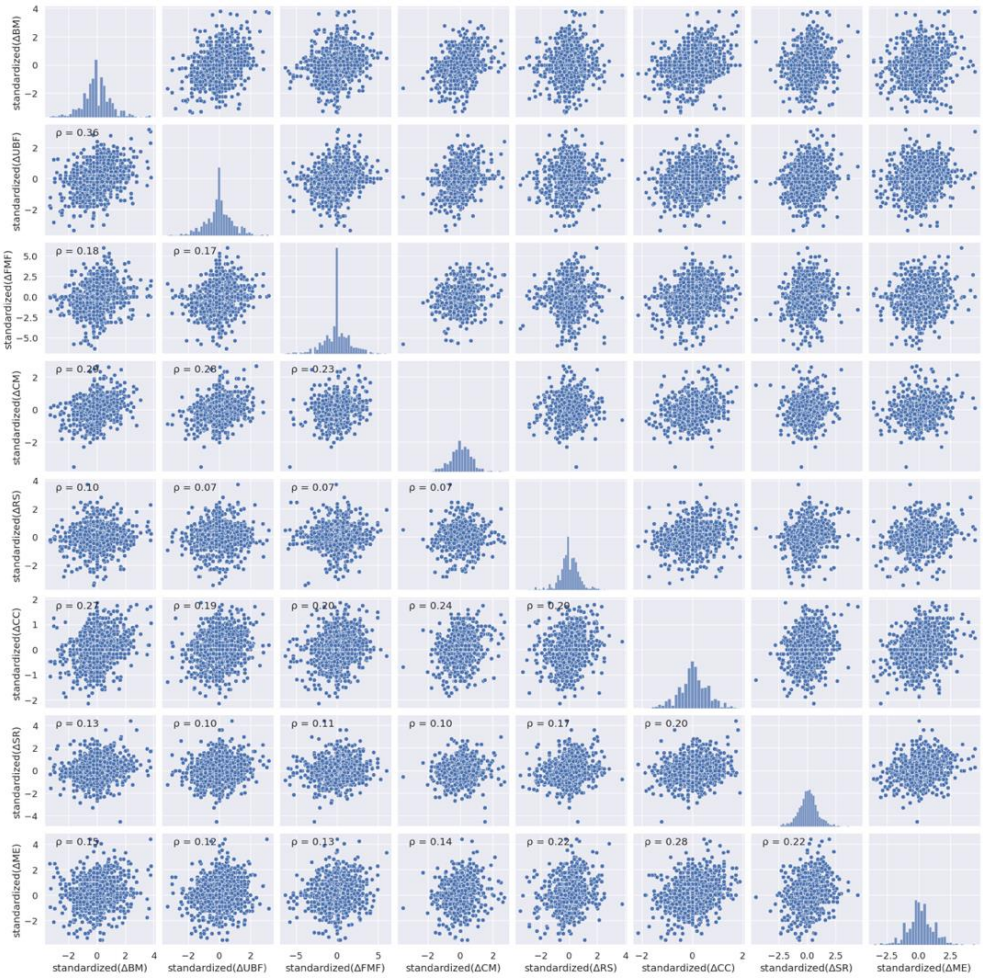

Figure 14. Scatterplot of Standardized Differences Between WD-FAB Administrations

For a 90% confidence interval, a standardized difference greater than 1.645 corresponds to a detectable change. Table 10 shows the proportion of the analytic sample that demonstrated change in each WD-FAB scale. From this table, we see that Fine Motor Function was most likely to have change in scores while Communication & Cognition was least likely to demonstrate change. Given the relatively short timeframe between the two administrations, it is not surprising that so few individuals experienced change in function.

Table 10. Detectable Change in WD-FAB Scale Scores (90% Confidence Interval)

	BM[4]	**UBF**	**FMF**	**CM**	**CC**	**SR**	**RS**	**ME**
SCORES IMPROVED	6.1%	3.3%	11.3%	1.4%	0.5%	4.1%	3.3%	7.1%
SCORES DECLINED	5.1%	3.3%	12.1%	0.8%	0.4%	3.1%	3.2%	4.6%

[4] BM: Basic Mobility; UBF: Upper Body Function; FMF: Fine Motor Function; CM: Community Mobility; CC: Communication & Cognition; SR: Self-Regulation; RS: Resilience & Sociability; ME: Mood & Emotions

Looking at individuals across all scales simultaneously and counting the number of scales where each individual improved or declined, Table 11 presents the observed frequency. At 90% confidence, approximately 72% of beneficiaries improved in no scales beyond the detection threshold, while the remaining 28% of participants improved beyond the threshold in one or more scales. Conversely, at 90% confidence, 74% of beneficiaries declined in no scales beyond the detection threshold. Given the relatively short timeframe between administrations, these results were consistent with our expectation.

Table 11. Number of Detectable Scale Score Changes (90% Confidence Interval)

	0	**1**	**2**	**3**	**4**	**5**
NUMBER IMPROVED	72.0%	21.1%	5.1%	1.4%	0.2%	0.1%
NUMBER DECLINED	74.1%	20.6%	4.3%	0.9%	0.2%	0.0%

When looking at overall improvement vs. decline, approximately 24% of participants improved in at least one scale score while declining in no scales (i.e., improved), 22% of participants declined in at least one scale score while improving in no scales (i.e., declined), and approximately 4% of participants had simultaneous improvement and decline in scales (i.e., mixed). Thus, half of the analytic sample had no detectable change in any scale (i.e., same).

When we consider the primary impairment type and corresponding functioning domain, the proportion of beneficiaries with no detectable change in any scale increases further (Table 12). Only 19% of beneficiaries with a physical primary impairment improved in one or more of the physical functioning scales, while approximately 17% of beneficiaries with a physical impairment declined in one or more physical functioning scales. About 12% of beneficiaries with a mental primary impairment improved in one or more mental functioning scales while 13% declined in one or more of the mental functioning scales.

Table 12. Number of Detectable Changes by Impairment Type (90% Confidence Interval)

		0	**1**	**2**	**3**	**4**
PHYSICAL	**Number Improved**	80.9%	17.3%	1.7%	0.1%	0.0%
	Number Declined	83.1%	15.3%	1.1%	0.5%	0.0%
MENTAL	**Number Improved**	87.8%	10.0%	1.5%	0.7%	0.0%
	Number Declined	86.7%	9.1%	1.7%	0.4%	0.0%

We next turn to the prediction of improvement in at least one scale score. Being able to predict improvement in WD-FAB scores is potentially relevant to CDR applications such as determining diary dates or refining the predictive modeling. To this end, we fit a Bayesian logistic regression model for improvement where we used the Survey 1 WD-FAB scores, demographic information, and SSA case attributes as additive linear predictors. The resulting standardized effects are shown in Figure 15 as log-odds ratios with mean and 95% posterior credible intervals. See Appendix C for a description of the variables and associated normalization factors used for standardization for the Bayesian models.

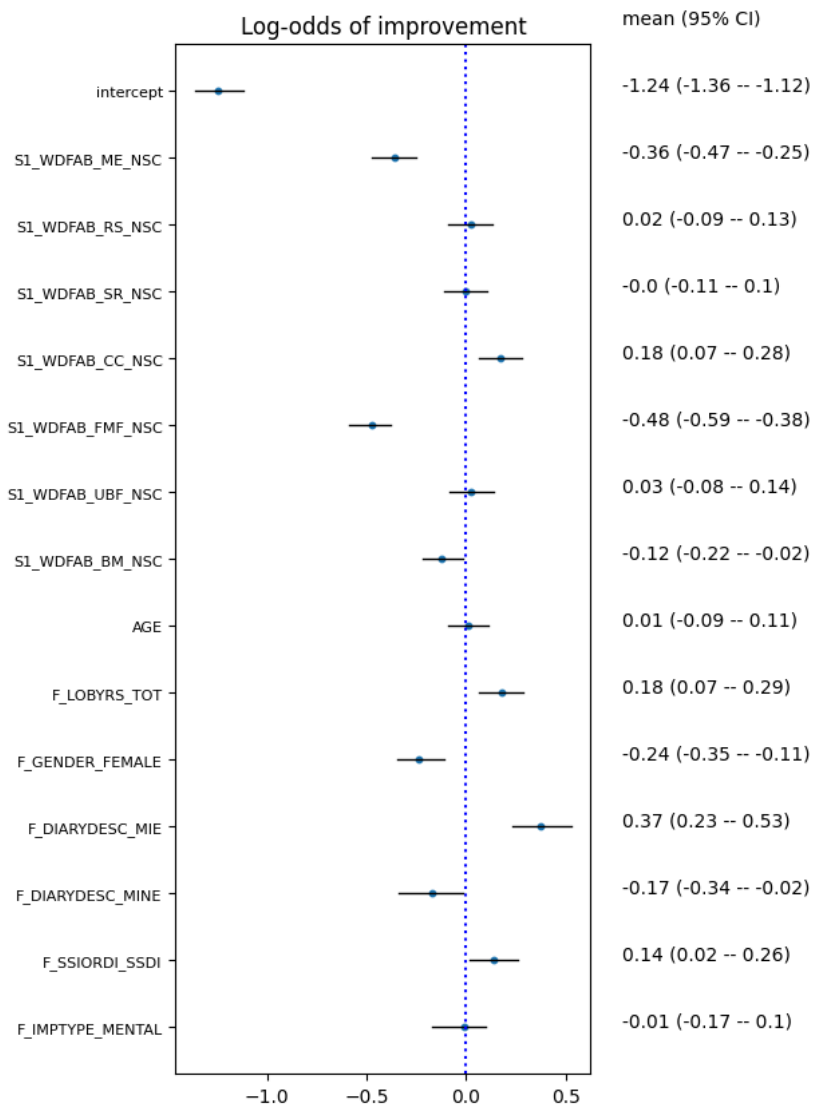

Figure 15. Log Odds of Improvement Beyond 90% Confidence Interval Threshold for Change

Individuals with lower baseline scores in Basic Mobility, Fine Motor Function, and Mood & Emotions had notably decreased odds of improvement beyond the 90% confidence interval threshold of change in at least one scale. Higher baseline Communication & Cognition scores were associated with increased odds of improvement in at least one scale. This model for predicting improvement scored an area under the receiver operator characteristic curve (AUROC) of 0.67, meaning the model correctly ranks the risk of approximately two-thirds of the paired comparisons between those who improved versus those who did not improve (Figure 16). The area under the precision recall curve (AUPRC) for this model was 0.409, in comparison to a frequency of 0.241 for the positive label. The literature generally places the metrics for this classification prediction in the upper-range of weak and the lower-range of acceptable.

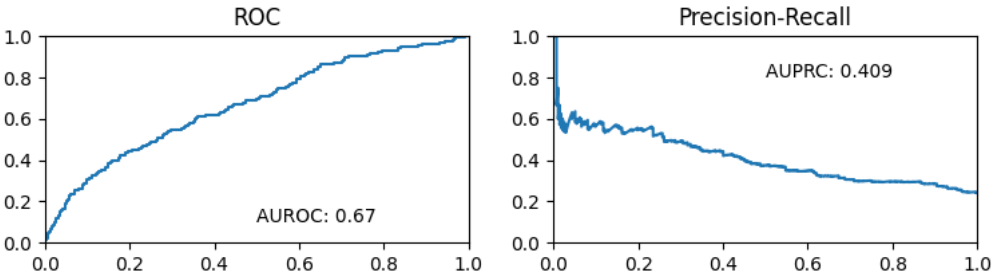

Figure 16. Receiver Operator and Precision-Recall Curves for Model Predicting Improvement

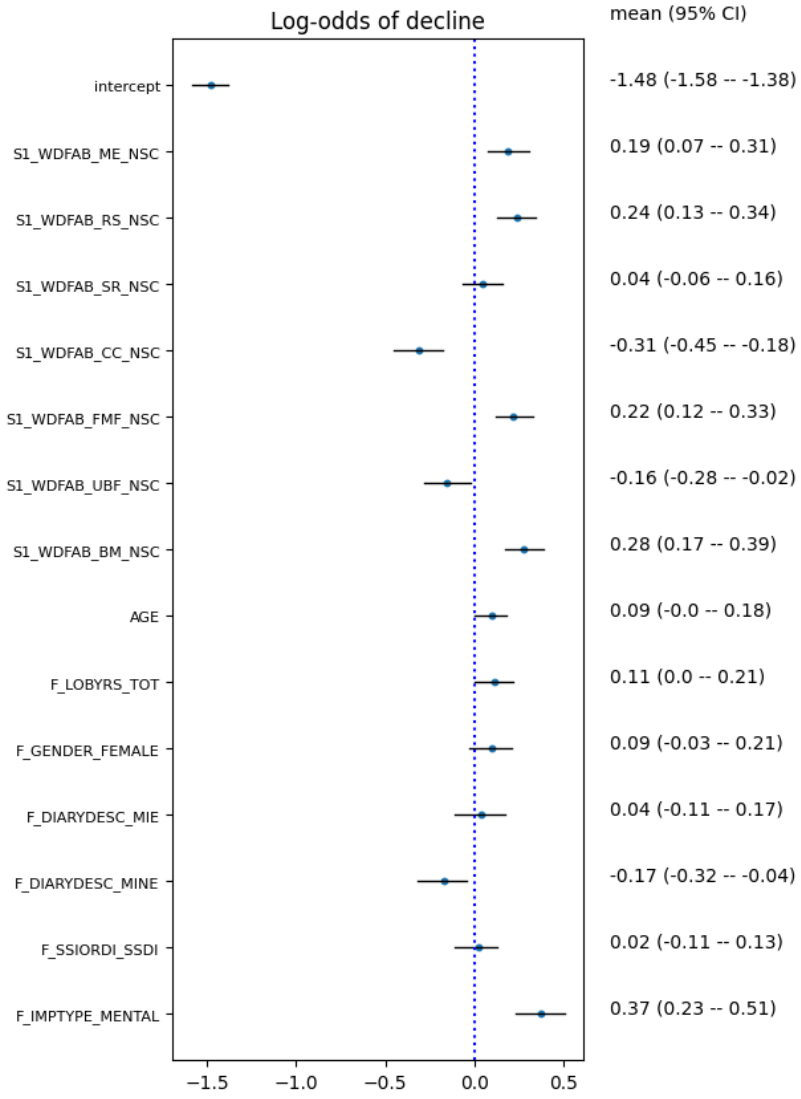

Figure 17. Log Odds of Decline Beyond 90% Confidence Interval Threshold for Change

We performed the same modeling to examine odds of decline beyond the 90% confidence interval threshold for score change. Figure 17 shows the standardized effects (in terms of log-odds ratios, mean and 95% posterior credible intervals shown). Those with higher Basic Mobility, Fine Motor Function, Mood & Emotions, and Resilience & Sociability scores at baseline were more likely to decline in at least one scale. Those with higher Upper Body Function and Communication & Cognition scores at baseline were less likely to decline in at least one scale. A mental primary impairment at CPD was also associated with increased odds of declining in at least one scale. The classification metrics for this model are presented in Figure 18 suggesting that the model is only weakly able to predict whether an individual will decline

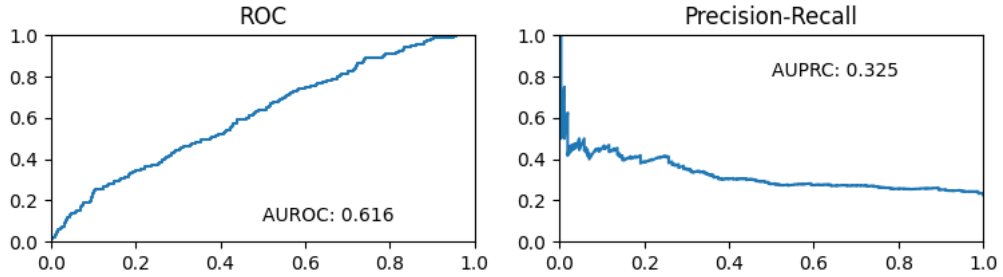

Figure 18. Receiver Operator and Precision-Recall Curves for Model Predicting Decline

In addition to looking at changes to the WD-FAB scale scores, we can also assess change in function using the WD-FAB functional profiles. If a beneficiary has a higher functional profile at Survey 2 than Survey 1, then we say there is functional improvement. If the functional profile is a lower category at Survey 2 than Survey 1, then this is decline. Having the same functional profile at Survey 1 and Survey 2 means there is no functional change. Note that because the functional profiles are defined using the functional levels within each scale, it is possible for a respondent to have a detectable change in WD-FAB scale score but not a change in functional level. Conversely, it is possible for a respondent to have a change in functional level that does not correspond to a detectable change in WD-FAB score if the scores sit close enough to the cut points between levels. Using the WD-FAB functional profiles to define change in function, we found that 19.9% of the analytic sample improved, 15.9% declined, and 64.2% remained the same.

Table 13. Demographics by Category of Change in Function

				PROGRAM TYPE			% PHYS	
SCORES	N	%	Mean Age	% SSI	% SSDI	% CONC	CPD	CDR
IMPROVE	300	24.1	47.9	58.0	38.3	3.7	63.0	65.0
DECLINE	275	22.1	47.2	56.7	37.8	5.5	56.4	58.2
SAME	621	49.9	47.1	63.0	34.1	2.9	66.7	67.6
MIXED	48	3.9	46.4	56.3	39.6	4.2	56.3	60.4
PROFILES								
IMPROVE	247	19.9	48.0	54.3	42.1	3.6	63.2	66.0
DECLINE	198	15.9	48.6	63.1	34.3	2.5	65.7	68.7
SAME	799	64.2	46.7	61.2	34.8	4.0	62.5	63.2

Table 13 provides demographic characteristics for the different sub-groups defined by change in WD-FAB scale scores and functional profiles. Notably, the proportion of beneficiaries with a physical primary impairment was lower for the sub-group that had a decline in WD-FAB scores compared to those who improved or had no change in scores. This trend was not observed in the changes to the WD-FAB functional profiles, though.

Comparing Change in WD-FAB Data to CDR Measures
As with the first study objective, we compared change in WD-FAB data to measures from across the various aspects of the CDR business process: CDR diary types, predictive model scores, mailer questions, and FMR evidence.

Figure 19 shows the distribution of change categories by CDR diary type using both WD-FAB scale scores and functional profiles. For both approaches to measuring change and across diary types, most beneficiaries did not demonstrate any change in functional ability. For beneficiaries with a diary of medical improvement expected (MIE), a higher proportion did improve across the two administrations when measuring change both with scores and functional profiles. When looking at changes in functional profiles, beneficiaries with a diary of medical improvement possible (MIP) also had a higher proportion of beneficiaries improve rather than decline.

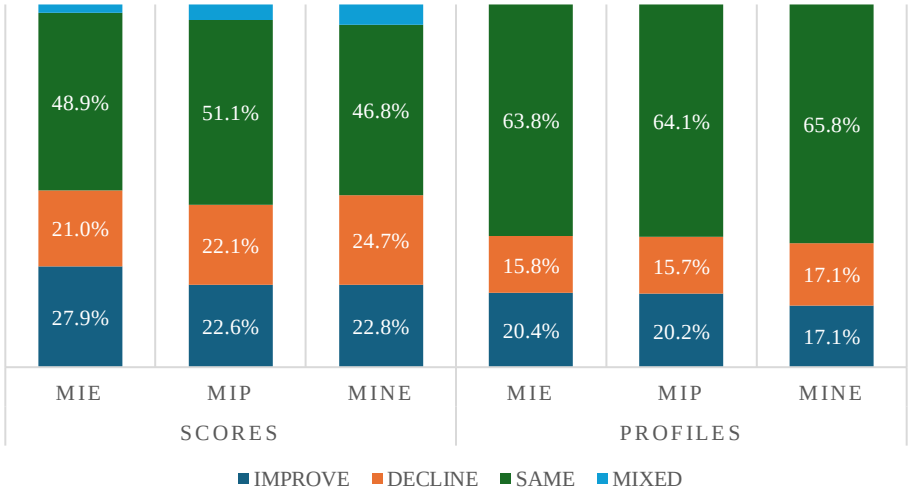

Figure 19. Distribution of Change in Function by CDR Diary Type (Scores vs. Profiles)

Figure 20 shows the distribution of change categories by CDR predictive model scores (PMS) using both WD-FAB scale scores and functional profiles. Similar to the diary types, most beneficiaries did not demonstrate change in function across the predictive model score categories whether measuring by change in score or functional profile. For beneficiaries with either a Medium or Low predictive model score, a higher proportion showed improvement than decline in both scores and functional profiles. Beneficiaries with a High predictive model score had a higher proportion of improvement than decline in the functional profiles.

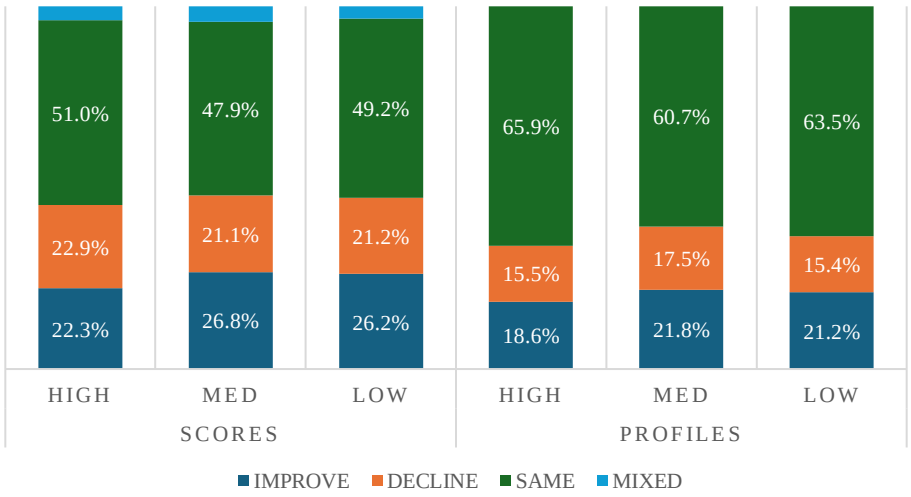

Figure 20. Distribution of Change in Function by Predictive Model Score (Scores vs. Profiles)

We also used logistic regression to model the relationship between changes in WD-FAB scores and the predictive model score. Significant results are included in Table 14. The predictive score is considered as an ordinal response variable. The odds of having a higher predictive model score are 1.07 times higher if the Survey 1 Basic Mobility score increases by 1 point. The odds of having a higher predictive model score are lower if the change in Mood & Emotions increases by 1 point.

Table 14. Logistic Regression Comparing Change in Function to Predictive Model Scores

Predictive Model Scores			
	Estimate	Std. Error	t-value
Survey 1 Basic Mobility	0.0678	0.01968	3.4452
Change in Mood & Emotions	-0.0326	0.01505	-2.1672
	Odds Ratio	Lower Bound	Upper Bound
Survey 1 Basic Mobility	1.0702	1.0300	1.1127
Change in Mood & Emotions	0.9679	0.9396	0.9968

We analyzed how the changes in WD-FAB scores relate to the responses to the CDR mailer responses from Survey 2. The results are included in Table 15. Higher Upper Body Function and Resilience & Sociability scores at Survey 1, as well as decline in Mood & Emotions scores were associated with a higher chance of having worked in the last 2 years (SSA445_1). The following factors led to a higher chance of having the same or better health over the past 2 years (SSA445_2): higher Basic Mobility, Upper Body Function, Fine Motor Function, and Mood & Emotions scores from Survey 1; improvement in Basic Mobility, Upper Body Function, and Mood & Emotions scores; lower Survey 1 Communication & Cognition scores; and decline in Communication & Cognition. Higher Survey 1 Upper Body Function and Resilience & Sociability scores, as well as improvement in Upper Body Function led to a higher chance of doctors' recommending return to work (SSA445_3). WD-FAB score changes were not

significantly associated with chance of hospitalization or surgery (SSA445_6), but lower scores in Basic Mobility and Mood & Emotions scales at Survey 1, higher Survey 1 Communication & Cognition scores, and decline in Basic Mobility led to a higher chance of having a doctor or clinic visit in the past 2 years (SSA445_7).

Table 15. Logistic Regression Comparing Change in WD-FAB to CDR Mailer (Survey 2)

SSA445_1: Worked in Last 2 Years				
	Estimate	Std. Error	t-value	Pr(>|t|)
Survey 1 Upper Body Function	0.0168	0.0048	3.4900	0.0005
Survey 1 Resilience & Sociability	0.0087	0.0034	2.5190	0.0120
Change in Mood & Emotions	-0.0070	0.0031	-2.2530	0.0246
SSA445_2: Health Last 2 Years				
	Estimate	Std. Error	t-value	Pr(>|t|)
Survey 1 Basic Mobility	0.0095	0.0043	2.2350	0.0258
Survey 1 Upper Body Function	0.0151	0.0052	2.8810	0.0041
Survey 1 Fine Motor Function	0.0072	0.0036	2.0180	0.0440
Survey 1 Communication & Cognition	-0.0113	0.0054	-2.0740	0.0385
Survey 1 Mood & Emotions	0.0141	0.0033	4.3170	0.0000
Change in Basic Mobility	0.0127	0.0048	2.6500	0.0082
Change in Upper Body Function	0.0136	0.0049	2.8090	0.0051
Change in Communication & Cognition	-0.0130	0.0060	-2.1530	0.0317
Change in Mood & Emotions	0.0116	0.0034	3.4500	0.0006
SSA445_3: Doctor Said Return to Work Last 2 Years				
	Estimate	Std. Error	t-value	Pr(>|t|)
Survey 1 Upper Body Function	0.0107	0.0037	2.8530	0.0045
Survey 1 Resilience & Sociability	0.0058	0.0027	2.1330	0.0333
Change in Upper Body Function	0.0075	0.0035	2.1480	0.0321
SSA445_7: Doctor/Clinic for Conditions Last 2 Years				
	Estimate	Std. Error	t-value	Pr(>|t|)
Survey 1 Basic Mobility	-0.0087	0.0022	-3.9920	0.0001
Survey 1 Communication & Cognition	0.0056	0.0028	2.0250	0.0433
Survey 1 Mood & Emotions	-0.0069	0.0017	-4.1380	0.0000
Change in Basic Mobility	-0.0064	0.0025	-2.6230	0.0089

We then looked at the relationship between the difference in WD-FAB scores and the difference in responses to the CDR mailer questions. The change between responses to the survey questions

was calculated by taking the difference between the Survey 2 and Survey 1 responses (i.e., Survey 2 minus Survey 1). The change variable for each survey question except SSA445_2 was considered as an ordinal variable with values of -1, 0, and 1. A value of -1 indicates a participant only has related activities in Survey 1, 0 indicates no change in responses, and 1 indicates a participant only has related activities in Survey 2. The change in SSA445_2 has values of -2, -1, 0, 1, 2, with a smaller value indicating improvement. Each question was considered as a dependent variable in the model. The predictor variables were WD-FAB scores in Survey 1 and change in WD-FAB scores from Survey 1 to Survey 2. The ordinal regression was run for the response variables. The regression models relate the predictor variables and the dependent variable and provide insights on how the changes in functional variables impact the changes in the survey responses. The results are shown in Table 16.

Table 16. Odds Ratios Comparing Change in WD-FAB with Change in CDR Mailer

SSA445_1: Worked in Last 2 Years			
	Odds Ratio	Lower Bound	Upper Bound
Survey 1 Self-Regulation	1.0765	1.0106	1.1468
SSA445_2: Health Last 2 Years			
	Odds Ratio	Lower Bound	Upper Bound
Survey 1 Communication & Cognition	1.0595	1.0107	1.1109
Change in Community Mobility	0.9669	0.9356	0.9993
Change in Communication & Cognition	1.0558	1.0014	1.1134
Change in Mood & Emotions	0.9685	0.9400	0.9979
SSA445_3: Doctor Said Return to Work Last 2 Years			
	Odds Ratio	Lower Bound	Upper Bound
Survey 1 Community Mobility	1.0598	1.0030	1.1200
SSA445_6: Hospitalized/Surgery Last 2 Years			
	Odds Ratio	Lower Bound	Upper Bound
Survey 1 Resilience & Sociability	1.0409	1.0005	1.0831
Change in Mood & Emotions	0.9613	0.9274	0.9963
SSA445_7: Doctor/Clinic for Conditions Last 2 Years			
	Odds Ratio	Lower Bound	Upper Bound
Survey 1 Mood & Emotions	0.9233	0.8686	0.9808

Beneficiaries with higher Self-Regulation scores at Survey 1 were more likely to have a positive change in reporting work status (SSA445_1). Those with higher Survey 1 Communication & Cognition scores, improvement in Communication & Cognition, as well as decline in Community Mobility and Mood & Emotions tended to have improvement in reported health status (SSA445_2). Beneficiaries with higher Survey 1 Community Mobility scores were more likely to have positive change in a doctor recommending they return to work (SSA445_3). Beneficiaries with higher Survey 1 Resilience & Sociability scores or decline in Mood &

Emotions were more likely to report an increased number of hospitalizations or surgeries between the two survey time points (SSA445_6). Finally, beneficiaries with lower Mood & Emotions scores at Survey 1 were more likely to report an increased number of doctor or clinic visits (SSA445_7).

When analyzing change in function, we also considered change in functioning information available between CPD and CDR. Starting at the case level, we calculated the proportion of cases with relevant functioning information for each WD-FAB scale at CPD and CDR, which provided the case presence per scale. We then took the difference between CPD and CDR (CDR minus CPD) as a measure of change in functioning information. Table 17 shows the change in case presence in comparison with the proportion of the analytic sample that demonstrated change in each WD-FAB scale to provide a sense of the relative amount of change in medical evidence compared with proportion of beneficiaries who experience change as measured by the WD-FAB. Only the scales of Upper Body Function and Self-Regulation had a higher proportion of cases with relevant functioning information available at the CDR time point rather than CPD. For all other scales, there was a higher proportion of cases at CPD with functioning information. While adjudicators are supposed to compare functioning information as part of the evaluation of signs and symptoms between CPD and CDR, this type of inconsistency in coverage of functioning information within the medical records may limit the adjudicator's ability to do so.

Table 17. Change in Case Presence from CPD to CDR Compared with WD-FAB Change

	BM[5]	UBF	FMF	CM	CC	SR	RS	ME
CHANGE IN CASE PRESENCE (%)	-2.88	7.74	-8.63	-8.85	-1.77	1.11	-0.44	-1.99
WDFAB SCORE IMPROVED (%)	6.10	3.30	11.30	1.40	0.50	4.10	3.30	7.10
WDFAB SCORE DECLINED (%)	5.10	3.30	12.10	0.80	0.40	3.10	3.20	4.60

At the mention level, we calculated the average amount of functioning information available in the medical evidence at CPD and CDR, as well as change in presence from CPD to CDR. Because of the differences in number and length of records between the two time points, we normalized the number of functioning mentions by taking the ratio of sentences with functioning information to sentences without functioning information, which we refer to as sentence normalization. We did the same calculation at the page level, which we refer to as page normalization. These calculations give measures of density that are comparable across cases.

Table 18 shows the average normalized presence at CPD, CDR, and change in presence from CPD to CDR as percent change (CDR presence divided by CPD presence) for all cases as well as by each category of change in function. Since CDR cases tend to include fewer records, the presence of functioning information is consistently denser in CDR cases with about 20% more mentions per sentence on average than in CPD cases. At the page level, we find that CPD cases have 0.57 mentions per page on average, compared to 0.64 mentions per page of CDR evidence,

[5] BM: Basic Mobility; UBF: Upper Body Function; FMF: Fine Motor Function; CM: Community Mobility; CC: Communication & Cognition; SR: Self-Regulation; RS: Resilience & Sociability; ME: Mood & Emotions

a 13% increase in density. The page-normalized counts provide an intuition of how sparse functioning information is in the medical evidence, occurring once every 1.5 to 2 pages on average. When looking across the different categories of change, the largest difference was for beneficiaries with mixed changes – improvement and decline in at least one scale score each – who had about 33% more mentions per sentence in the CDR evidence on average. For beneficiaries whose function improved over the two administrations, the density of the functioning information still increased from CPD to CDR, but to a much smaller degree than the average case.

Table 18. Presence of Functioning Information as Average Mentions per Sentence and per Page by WD-FAB Change in Function Category

	AVERAGE CPD PRESENCE		**AVERAGE CDR PRESENCE**		**CHANGE IN PRESENCE (%)**	
	Sentence	*Page*	*Sentence*	*Page*	*Sentence*	*Page*
ALL CASES	0.0071	0.57	0.0086	0.64	20.2	12.7
SCORES						
IMPROVE	0.0068	0.53	0.0077	0.57	13.2	6.5
DECLINE	0.008	0.64	0.0095	0.72	19.3	11.4
SAME	0.0069	0.54	0.0085	0.62	22.9	15.5
MIXED	0.0054	0.44	0.0072	0.56	33.4	27.0
PROFILES						
IMPROVE	0.0081	0.64	0.0087	0.64	7.5	0.5
DECLINE	0.007	0.57	0.0084	0.63	16.9	11.5
SAME	0.0068	0.54	0.0086	0.64	25.6	17.3

Relating Changes in WD-FAB Data to CDR Outcomes

For our final set of analyses, we looked at how CDR FMR outcomes relate to changes in WD-FAB data. Table 19 shows the FMR initial allowance rate (% CO) for each of the sub-groups corresponding to change in WD-FAB scores or profiles.

Table 19. CDR Initial Allowance Rate by Change Sub-Group

	N	**% CO**
TOTAL	1244	93.4
SCORES		
IMPROVE	300	95.0
DECLINE	275	92.4
SAME	621	93.1
MIXED	48	93.8
PROFILES		
IMPROVE	247	94.7
DECLINE	198	90.4
SAME	799	93.7

We fit a Bayesian logistic regression model to the task of predicting the outcome of cessation (dcn_cd = CE). Figure 21 shows standardized effects measured as log-odds ratios with mean and 95% posterior credible intervals shown. Having a mental primary impairment at CPD was associated with the largest decrease in the odds of cessation. WD-FAB data (both baseline scores and change) generally had only small associations with a cessation decision, although beneficiaries who had a decline in their functional profile had higher odds of cessation. The largest association was with the Upper Body Function scale; those with higher initial Upper Body Function scores and those with larger improvements in Upper Body Function scores were most likely to have a cessation. This model had good predictive accuracy as measured by the AUROC as shown in Figure 22.

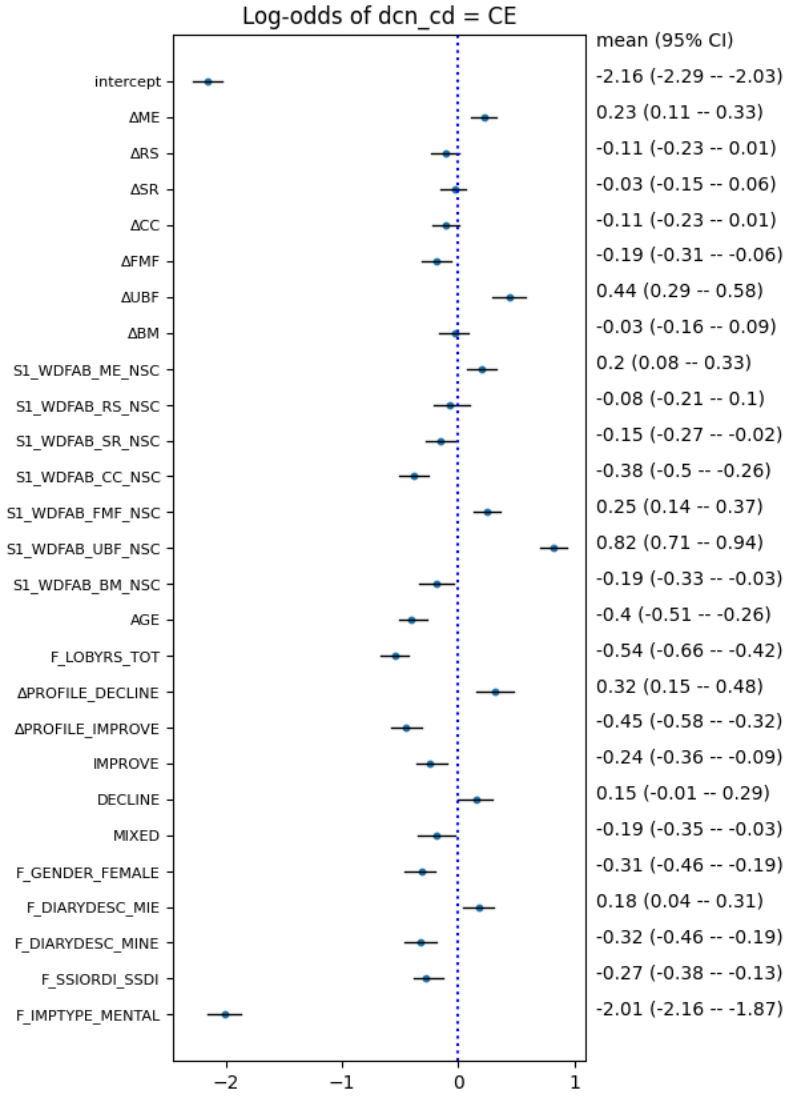

Figure 21. Log Odds for Predicting Cessation Cases (Mean and 95% Confidence Interval)

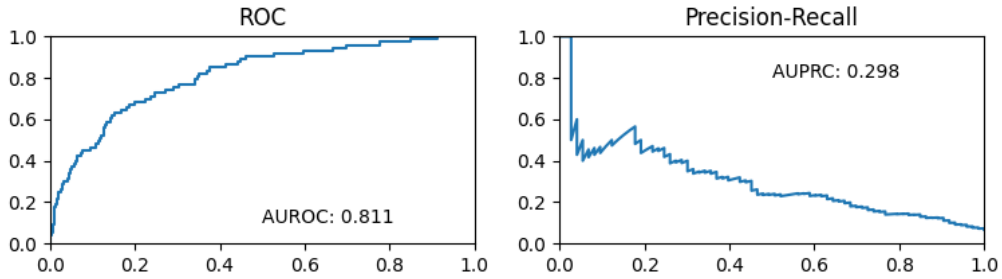

Figure 22. Receiver Operator and Precision-Recall Curves for Model Predicting Cessation

We also considered the relationship among change in WD-FAB data, change in medical evidence, and the FMR outcomes. For the medical evidence, we looked at the ratio of cases with functioning information at CPD vs. CDR for continuances vs. cessations. This was calculated for all cases as well as by CDR and WD-FAB category. Cessations tended to have a higher proportion of cases with more functioning information available at CDR than CPD. However, when looking at the amount of functioning information using the sentence normalization, cessation cases tended to have fewer functioning mentions on average. Figure 23 shows the average presence of functioning information for continuances vs. cessations for all cases as well as for the CDR predictive model score (PMS) categories, the first two CDR mailer questions pertaining to work and health status, as well as change in function measured using WD-FAB scores and functional profiles. While continuances tend to have more functioning information on average for all cases and most sub-categories, cases where there was functional improvement, measured either by change in WD-FAB scores or profiles, tended to have more functioning information present for cessations rather than continuances. This might suggest that there was more documentation of this functional change in the record to support the decision of medical improvement.

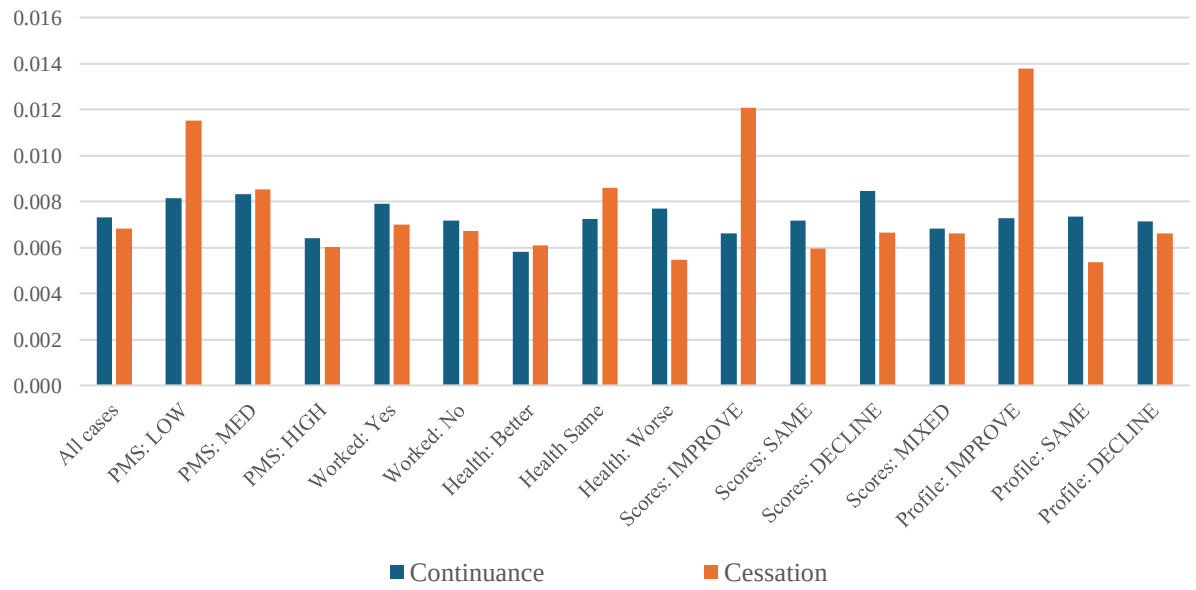

Figure 23. Presence of Functioning Information in Medical Evidence for Continuances vs. Cessations

These analyses on the medical evidence excluded the disability determination explanation (DDE), as this document is generated by the adjudicator as a summary of the case and is not evidence available when reviewing the case. However, the DDE provides insight into the evidence requirements an adjudicator relies on to support their decision, especially when demonstrating that there has been medical improvement related to the ability to work (i.e., applying the MIRS for cessation). Therefore, we conducted a separate analysis with the DDE files looking at presence of functioning information and how this information compares with the WD-FAB and other CDR measures. The DDE files were denser in functioning information than the medical evidence. When comparing by FMR outcome, DDEs for cessations were longer than for continuances with an average of 563 vs 413 sentences, respectively. Overall, DDEs for both continuances and cessations had a similar density of functioning information. When we looked at sub-categories based on CDR and WD-FAB values, though, we saw certain trends in the relationship of density of functioning information in the DDE for cessations vs. continuances.

Figure 24 shows the ratio of average presence of functioning information in the DDE for cessations over continuances for all cases (in black) vs. CDR categories (in green) vs. WD-FAB categories (in light blue). Values greater than 1.0 indicate that there was more functioning information in the DDE on average for cessations than for continuances, and conversely, values below 1.0 indicate that the DDE for continuances included more functioning information on average.

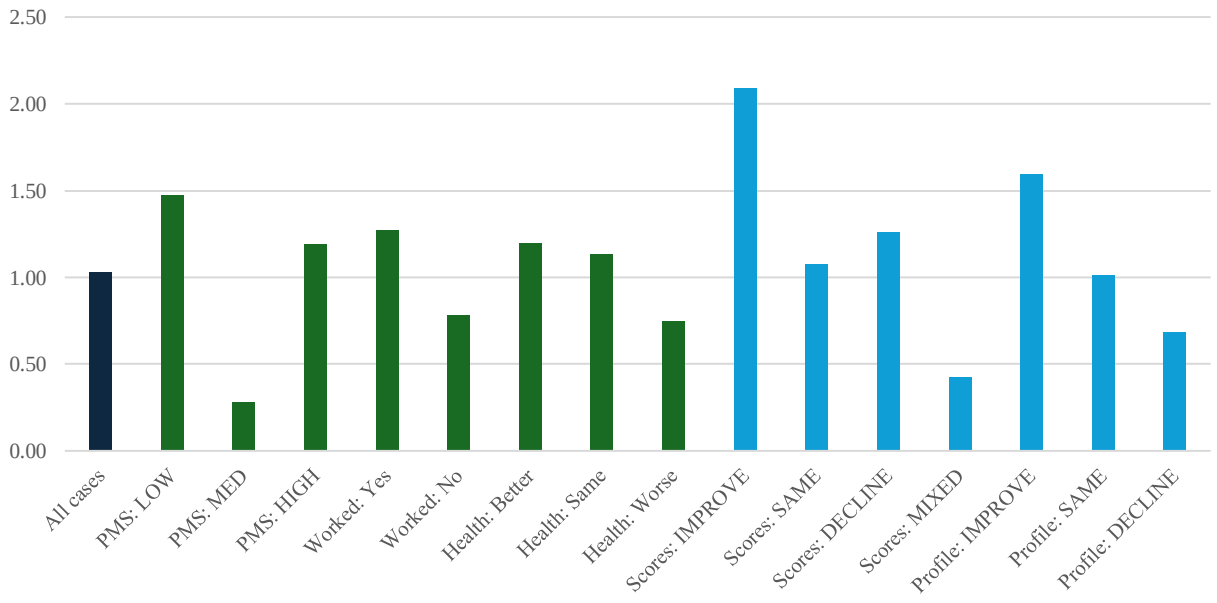

Figure 24. Ratios of Presence of Functioning Information in the Disability Determination Explanation for Cessations over Continuances

For the CDR mailer questions related to employment and health status, beneficiaries who had been employed or had better health, characteristics associated with improvement, had a higher ratio of functioning information in the DDE for cessations than continuances than beneficiaries who had not worked or had worse health. A similar trend is observed in the WD-FAB data where beneficiaries who had functional improvement, measured by changes in either WD-FAB scores

or functional profiles, had a higher ratio of functioning information in the DDE for cessations than continuances than beneficiaries whose function had declined. Thus, the WD-FAB, as a measure of functional change over time, has a similar relationship to evidence cited for CDR decision making as existing CDR measures while having the added advantage of being more standardized, comprehensive, and easier to compare over multiple administrations.

Value and Feasibility of Integrating the WD-FAB into the CDR Process

Measuring Respondent Burden

The WD-FAB Research Study offered the first opportunity to collect data from the beneficiary population. Previous studies with the WD-FAB had focused on claimant or general working age adult populations. Thus, as part of looking at a new population and the context of the CDR process, we wanted to assess whether administration metrics would remain consistent and whether there was respondent burden that might affect the impact of the WD-FAB in the CDR context. We used both quantitative and qualitative methods to measure respondent burden. For each WD-FAB administration, data on the number of items fielded and the time to complete each scale as well as overall administration time was collected. We found that mean administration times were consistent with previous studies (Table 20). Administration time was longer for those who responded via telephone, which likely relates to the time required to read the questions aloud. However, there were more outliers in the web administration since respondents could take breaks and come back to complete the survey on their own time, which is consistent with previous work examining web-based administration. Fine Motor Function tended to take the least amount of administration time (1.19-1.26 minutes on average) while Communication & Cognition took the longest (2.01-2.17 minutes on average). Additional administration metrics for each WD-FAB scale are included in Tables D1 and D2 in Appendix D.

Table 20. Administration Metrics by Survey Time Point and Mode

		Mean Administration Time (min)	Mean Total # of Items	Mean Time per Item (sec)
Survey 1	Total	14.14	70.85	11.97
	Telephone	17.63	69.01	15.33
	Web	12.70	71.56	10.65
Survey 2	Total	13.32	70.75	11.30
	Telephone	17.38	67.91	15.35
	Web	12.04	71.74	10.07

As part of the first survey administration, respondents were asked four feedback questions about whether they found responding to the survey to be burdensome, confusing, or difficult, and whether they were comfortable answering the questions. Respondents were also given the opportunity to provide general feedback in addition to these four questions. These questions were meant to correspond just to the WD-FAB; however, because these questions were administered at the end of the survey, which included the WD-FAB along with the CDR mailer questions, demographics, and health status questions, some respondents reflected on their experience of the entire survey rather than just the WD-FAB portion. In addition, the original study design called for these questions to be administered as Likert items with categorical response options, but

instead the survey was designed using free text responses. Thus, analyzing the collected feedback required qualitative methods. To process the free text data, we removed extra spaces and special characters to make responses more uniform, and then we classified responses into Likert-type categories (e.g., a little, some, a lot, etc.) to reflect the original study design. For the fifth question that asked respondents to provide any additional feedback, responses were classified as positive, negative, mixed feedback, or general comments.

These data were collected as part of the first survey; therefore, the results come from the larger sample of 2,406 beneficiaries who completed this administration. Figure 25-Figure 28 show the distribution of responses for the four feedback questions. Across all four questions, the majority of respondents (at least 65%) expressed having positive experiences responding to the survey. Approximately 68% reported that answering these questions was not burdensome (Figure 25); nearly 70% felt the questions were not confusing (Figure 26); 65.4% had no difficulty answering these questions (Figure 27); and 72.6% were comfortable answering the questions (Figure 28). An additional 25% to 32% of the respondents experienced various levels of response burden, ranging from a little to a lot or from slightly disagree to strongly agree. In these figures, the category of "Other" includes those who did not respond, responded don't know, or the response was not applicable to the question.

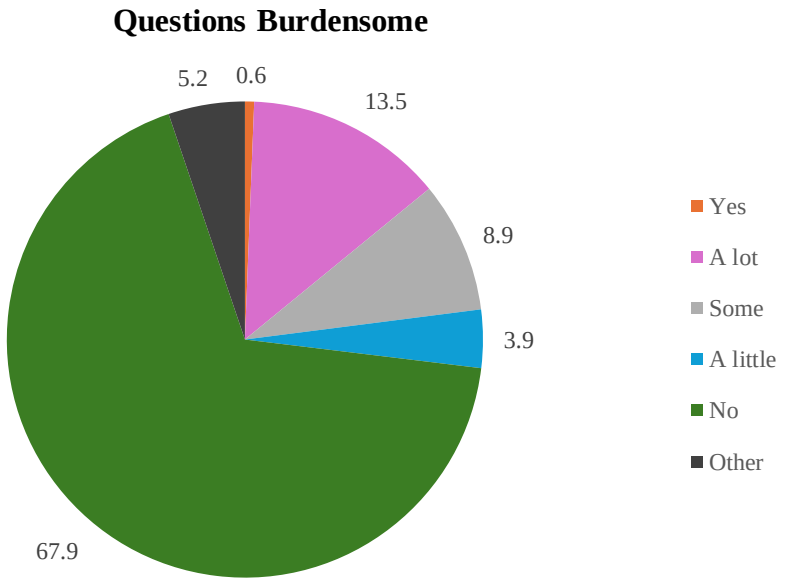

Figure 25. Pie chart of Likert-type Responses for whether Questions were Burdensome

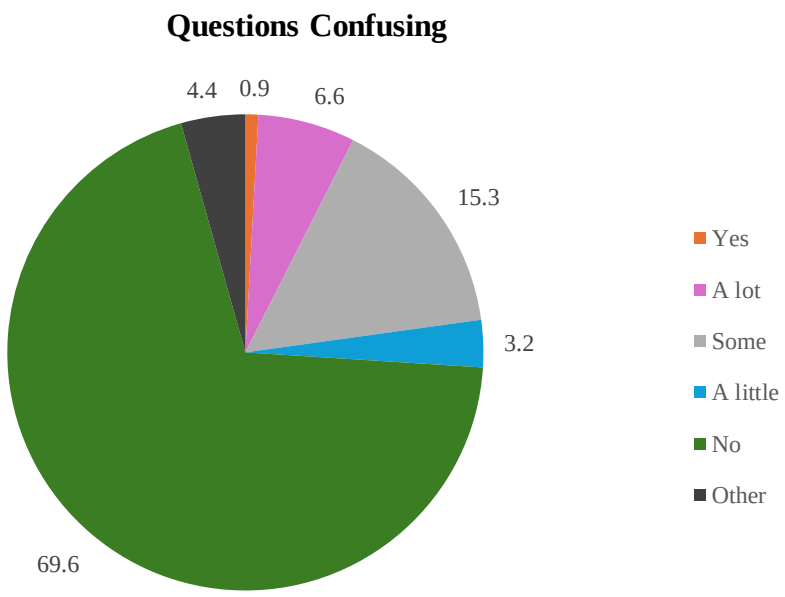

Figure 26. Pie chart of Likert-type Responses for whether Questions were Confusing

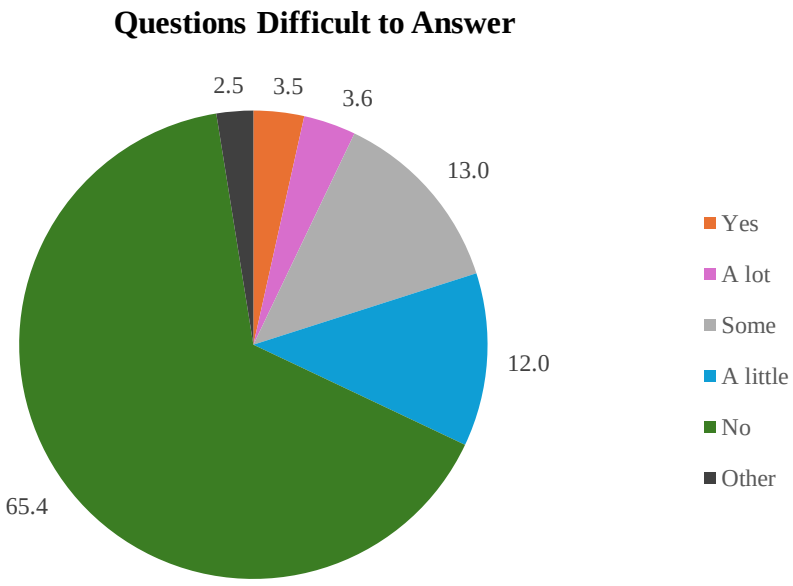

Figure 27. Pie chart of Likert-type Responses for whether Questions were Difficult

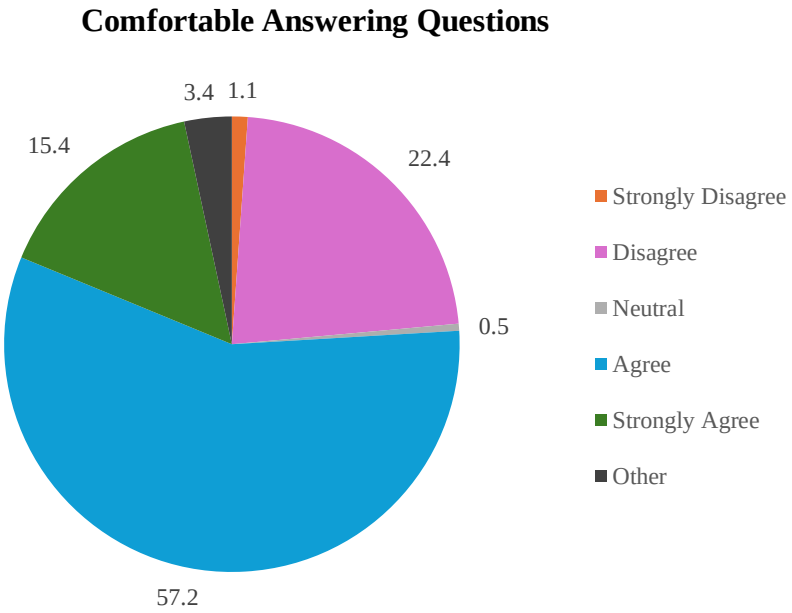

Figure 28. Pie chart of Likert-type Responses for whether Comfortable Answering Questions

For the fifth question where respondents could provide additional feedback, 560 (23.3%) respondents chose to include comments. These responses included both positive and negative feedback on the survey experience, as well as mixed feedback and general comments. Examples of positive feedback include that the questions were valid, fair, interesting, concise, easy to understand, straightforward, well thought out, thorough but not intrusive, and easy to answer. For the negative feedback, impressions included that the questions were repetitive, too broad (not disability specific), too personal, difficult to understand, and difficult to remember past events (e.g., doctor's appointments). Individuals with mental and sensory disabilities felt that questions specifically related to their disabilities were not available. Some respondents felt stressed or uncomfortable answering the questions because they did not fully understand the purpose of the survey and were afraid that their responses would affect their disability benefits. Because of health or disability related issues (e.g., having difficulty staying focused), respondents also expressed the need for personal assistance or taking multiple breaks to complete the survey. For a small number of respondents, English was not their native language, and so the language barrier was another source of burden.

For some, the modality of answering the questions seemed attributable to varying levels of response burden. This was also relative to the nature of the underlying health conditions the respondents reported having as well. For instance, manually entering responses online was challenging and time consuming for people with physical disabilities. A phone interview was difficult for people with hearing loss (e.g., hard of hearing). Due to time and memory constraints, some respondents found it difficult to answer questions with multiple choices on the phone. In addition, respondents had different opinions about response choices. Some preferred binary questions than multiple-choice questions. Others felt that additional response choices were needed because current options did not sufficiently or adequately reflect their level of functioning. Some respondents also indicated that their level of functioning could vary due to

different circumstances (e.g., health related issues, time of the day, or locations). This suggests that having repeated administrations of the WD-FAB would better represent the beneficiary population and support SSA's business process.

Conducting Focus Groups on the Potential Value of the WD-FAB in the CDR Process
Since the WD-FAB data were not collected as part of the CDR process itself for this pilot study, preventing us from directly measuring impact, we leveraged focus group meetings with CDR stakeholders to collect feedback on potential applications of the WD-FAB in the CDR process and initial impressions on benefits of the instrument, as well as areas that would require additional resources or training in order to use the WD-FAB effectively. We held three focus group sessions with SSA employees from DCO/ODD, OARO/OQR, ODP/OMP, OPSOS, as well as the DDS between May 2022 and March 2024. The main aims of the first focus group session were to provide the NIH team with background on the CDR process and introduce the WD-FAB to the SSA participants for initial reactions and thoughts around use in the CDR process. The second focus group was held after the first survey data collection and focused on the CDR mailer process as well as the potential use of single administration WD-FAB data. The third and final focus group session was held after data from both survey administrations were available and looked at case studies focused on change in WD-FAB data to get feedback from SSA employees involved in the full medical review process. For each focus group, the session was recorded and transcribed using MS Teams' meeting transcription feature. The transcripts were then reviewed and coded for the type of feedback provided. Across the focus groups, feedback on the WD-FAB was grouped into the following four categories:

Application: how the WD-FAB could be used;
Benefit: potential value or a positive feature of the WD-FAB;
Limitation: potential weakness or a problematic feature of the WD-FAB;
Training: what is needed to use the WD-FAB correctly.

Each category included a number of common themes that were expressed in one or more of the focus group sessions. Table 21 lists the common themes that aligned with each category along with specific examples of feedback that corresponded to the themes from across the focus group sessions.

Table 21. Focus Group Feedback Categories, Themes, and Specific Examples

CATEGORY	THEME	EXAMPLE
APPLICATION	• When WD-FAB could be administered in the CDR process • Use for assessing new impairments if there is medical improvement • Obtaining information usually needed in the review process • Reference tool for training new adjudicators	*"… would be very good at that field office level to decide if we need a CDR when we get to the DDS side of things."* *"… would like some of the information to somehow be integrated into that 454 and that information they give us because that 454 information is very dry and it really doesn't give you much"*

		"… role similar to an ADL form or … the ADL section on the CDR report where you might administer the WD-FAB and then have a follow up question if they would like to explain any of their answers…"
BENEFIT	• Relevant and/or specific content • Content administered consistently within and across respondents and timepoints • WD-FAB is a useful tool • Collects comprehensive information on function across domains • Provides better quality data	*"… able to kind of get cleaner, better quality data when it comes to functioning"* *"…will help guide where we need to investigate"* *"This is actually enforcing what the predictive model says and gives it a better push to know what we do have to look at the DDS level"* *"…because it is very valuable information and what's really great about the way these questions are and the way that the person is able to answer them is it helps them articulate something they might not be able to articulate on their own."* *"…provides a way to have some standardization in what tasks they're being asked about"* *"… can compare self-reported function from one period to another"*
LIMITATION	• Not able to see and/or compare the same questions across administrations • All respondents do not receive exact same questions • Potential burden of responding or resistance to responding • Potential for WD-FAB to over influence direction of case development	*"…what we're all talking about is self-reported function, which is only a small part of the whole process"* *"The fact that there's 300 questions, but we might only see 60 and it's a different 60 for each individual person is what would give me pause as to this is what it is and we wouldn't use like number scatters or graphs while that's good information, we would actually have to have that objective answer."* *"A 300 questionnaire or, you know, even at 60 or 70 questions, people will become resistant to answering all of*

		those questions just because that's the nature of people, unfortunately. "
TRAINING	• Proper use of scores and not overweighting objective scores • Interpreting WD-FAB scores as part of CDR cases • How individual questions inform scale scores • How to use WD-FAB in CDR business process	*"…if you could show that there's a standardized methodology to it and this is very consistent, then it would [work], it would not be an easy haul."* *"I do think that would be helpful to like when you report that kind of zone in on the score led to this and now pay attention to this."* *"… would we be able to have the ability to compare this to the objective, more objective findings on file to determine how supportable their self-report is?"* *"… if you have something like that [low score], does that rise to the threshold of a potential impairment that has to be developed…"*

All four categories were relevant and sufficient for analyzing the feedback received across the three focus group sessions. At each session, there was enthusiasm and excitement around the potential use of the WD-FAB for the structure and consistency it would bring to collecting data on function. While the focus group participants did raise concerns, many of these were due to a lack of familiarity with the instrument that can be addressed through the provision of additional resources and training around the WD-FAB as a measurement tool, the scores, and how to use these data as part of SSA business processes. Some of these resources already exist such as brief write ups on the WD-FAB, item maps for interpreting scores, and a user guide, which would have to be tailored to the SSA use cases and adjudicators. Training on using the WD-FAB as part of the CDR process would need to come from SSA, though.

The feedback raised during these sessions is helpful for exploring options for introducing the WD-FAB into SSA's CDR business process and the type of information about the instrument that would be needed to make it useful to SSA employees, which are reflected in the summary recommendations in the subsequent section.

Summary Recommendations

We summarize our recommendations on the potential role of the WD-FAB in the CDR business process by considering how the findings from this pilot study support three different use cases or applications.

Application 1: Incorporating WD-FAB Scores into the CDR Predictive Models
The WD-FAB can be used as part of the screening process to help SSA determine who should undergo a FMR. From these analyses, we see associations between baseline WD-FAB scale scores as well as changes in scores and the likelihood of cessation such that including WD-FAB data as part of the current predictive models has the potential to refine the model estimates further. As part of previous work analyzing WD-FAB data from the Supported Employment Demonstration (SED), which included annual WD-FAB data over a three-year period, we see that similar results hold for predicting employment outcomes [11]. While the SED focused on a different study population, findings were consistent across the three study years suggesting that effects observed as part of this pilot study might also hold or become stronger with repeated administrations over time.

SSA currently runs the predictive models annually for all beneficiaries and then looks at scores for those beneficiaries whose CDR diaries have come due. Therefore, in order to be able to incorporate WD-FAB data into the predictive models, we recommend collecting WD-FAB from all beneficiaries annually. With an annual data collection, both the baseline WD-FAB scores as well as the year-over-year change in scores could be used as factors in the predictive models. This would provide SSA with more regular insight into beneficiaries' status and how their function changes over time. Having access to such data also offers an opportunity to potentially refine the diary types and how diary dates get assigned in the future. While directly measuring change over time would be most effective, we have shown with these data that likelihood of improvement can also be predicted, although the performance of this model was not strong.

Incorporating WD-FAB data into the CDR predictive models would require having the WD-FAB scores stored in a database that would be accessible to the systems currently used to run the predictive models. Training on the WD-FAB would focus on more of the technical requirements needed to use these data as model parameters.

Application 2. Using WD-FAB Scores in Addition to or Instead of the CDR Mailer
The WD-FAB can be collected as part of, or instead of, Form SSA-455 (the CDR Mailer). When comparing WD-FAB scores to the CDR mailer responses, we found relationships between both the physical and mental functioning scales and likelihood of having been employed, health status, and healthcare use. The relationships between the WD-FAB scores for Upper Body Function and Resilience & Sociability and employment status are again consistent with results from the SED over a three-year period. In addition to demonstrating relationships with the existing CDR mailer questions, the WD-FAB has the added advantages of providing more consistent data, being comparable both across respondents and across administrations, and offering greater insight into functional abilities through score interpretation. Leveraging WD-FAB data also promotes evidence-based decision making that can look beyond just the completeness of the CDR mailer responses to consider both current level of functioning and whether a beneficiary has demonstrated sufficient change in function to be likely to earn above SGA and therefore warrant spending resources to conduct a FMR. This pilot study demonstrated that change in function can be measured through both the WD-FAB scale scores as well as the functional profiles. Further work would be needed for SSA to determine which measure of change is more relevant to the business process. In either case, these WD-FAB data offer a standardized measure for defining when change has, or has not, occurred.

If the WD-FAB is collected annually to inform the predictive models, then there would not be a need for a separate data collection as part of the CDR Mailer. If WD-FAB data are not collected annually, then the WD-FAB could be administered just to those beneficiaries whose diary dates have come due. SSA would then have the current WD-FAB scores as well as change in WD-FAB scores from any previous administrations, such as WD-FAB data collected at the CPD, to inform CDR decisions.

Application 3. Assessing Function as Part of the Full Medical Review
WD-FAB data can serve multiple purposes as part of the full medical review. The WD-FAB can be used to supplement existing processes such as refining or replacing the daily activities section of Form SSA-454 (CDR Report). All eight WD-FAB scales demonstrated expected relationships with the current SSA-454 Daily Activities with lower WD-FAB scores, which represents lower functional abilities, corresponding to greater difficulties with the daily activities reported on average. The WD-FAB has the added advantages of covering a wider range of activities and providing greater granularity beyond the binary 'Yes' or 'No' of the current form.

The WD-FAB can also be leveraged to assist with case development by identifying areas of limitation that should be investigated and developed within the medical evidence. By comparing with the NLP output, we showed that WD-FAB data can be associated with findings from the medical records such that decisions can still be made on the basis of medical evidence. We also observed similar patterns in the amount of functioning information available in the medical records and used to support the disability determination across existing CDR measures and the WD-FAB suggesting that the WD-FAB can be just as relevant as other tools used within the CDR process while having the advantage of rigorous development specifically for SSA's use and client population.

Since the FMR is driven by the MIRS, WD-FAB data can also be used to measure functional change as part of evaluating signs and symptoms from the CPD. However, under current SSA business regulations, in order for WD-FAB data to be an admissible measure to be included as part of the side-by-side comparison between CPD and CDR, the WD-FAB would have to be included as a measurement as part of the CPD.

Conclusion

The data and findings from this WD-FAB Research Study establish a foundation of evidence for assessing the potential application and utility of the WD-FAB as part of SSA's CDR business processes. Results demonstrated that the WD-FAB has the expected relationships with similar measures of function (SSA-454 Daily Activities) while providing new insight into how the WD-FAB aligns with and can be used in the Continuing Disability Review process.

As part of the study analyses, we examined single administration WD-FAB data as well as change in WD-FAB data in relation to CDR data and outcomes, and we also assessed the value and feasibility of using the WD-FAB in the CDR process. From single administration data, WD-FAB scores of beneficiaries indicated limitations in both physical and mental functioning that were consistent with other self-reported limitations such as difficulties with daily activities documented in Form SSA-454, the CDR Report. WD-FAB scores were also related to CDR mailer responses and supported by functioning information identified from the medical evidence.

When analyzing change in WD-FAB data, the majority of beneficiaries did not demonstrate functional change, which is to be expected over a six-month timeframe based on previous literature [12]. There was some functional change observed, though, defined both through standardized score differences and shifts in functional profiles. While these changes were analyzed in comparison to CDR case decisions, the timing of the data collection in relation to FMR decisions was not consistent across the cases, limiting any conclusions about relationships between WD-FAB data and case outcomes. Despite these limitations, we did see trends in WD-FAB data that were consistent with observations from existing CDR measures, such as the CDR predictive model scores and mailer responses, suggesting that the WD-FAB could be a relevant and complementary data source in the CDR process.

Administration metrics demonstrated that the WD-FAB is a highly efficient measure of functional ability with scores generated across the eight dimensions in under 15 minutes on average. While the majority of study participants opted to take the WD-FAB over the internet, there are notable advantages to having multiple administration modes available to respondents as indicated by differences in demographics and feedback provided from beneficiaries. Having multiple administration modes available to respondents means beneficiaries could take the WD-FAB from the comfort of their own homes or during a provider visit for those who do not have internet access, or it could be administered via phone as an alternative mode that might be preferred by some demographics. Each mode will likely benefit different sub-populations with different needs or health issues. One observation from the feedback that SSA would need to address is having a way to administer the WD-FAB for beneficiaries whose primary language is not English. We would recommend having a Spanish translation of the instrument at a minimum.

Additional work is needed to sufficiently justify and support the integration of the WD-FAB as a complementary information source. Namely, the refinement and development of resources to help adjudicators understand and interpret WD-FAB data in the SSA context, as well as additional studies that include the WD-FAB as part of actual CDR cases to assess impact on decision making. These preliminary findings demonstrate that there is potential value and benefit for SSA and the CDR business process in pursuing these next steps.

References

1. 42 U.S.C § 421(i) (2015)
2. Social Security Administration. *Annual Report on Medical Continuing Disability Reviews Fiscal Year 2014*. December 2016. Online. https://www.ssa.gov/legislation/FY%202014%20CDR%20Report.pdf
3. Social Security Administration. "DI 28001.020 Frequency of Continuing Disability Reviews (CDRs)." *Program Operations Manual System (POMS).* 2023. url: https://secure.ssa.gov/poms.nsf/lnx/0428001020
4. Social Security Administration. "Social Security Administration (SSA) Annual Data for Periodic Continuing Disability Reviews (CDR) Processed and Backlog." *Open Government Initiative*. 2023. url: https://www.ssa.gov/open/data/Periodic-Continuing-Disability-Reviews.html
5. Social Security Administration. *FY 2025 President's Budget*. March 2024. Online. https://www.ssa.gov/budget/assets/materials/2025/2025BO.pdf
6. Social Security Administration. "DI 28010.015 Comparison of Symptoms, Signs, and Laboratory Findings When Evaluating Medical Improvement (MI)." *Program Operations Manual System (POMS).* 2023. url: https://secure.ssa.gov/poms.nsf/lnx/0428010015
7. Marfeo EE, Ni P, Chan L, Rasch EK, McDonough CM, Brandt DE, Bogusz K, Jette AM. Interpreting physical and behavioral health scores from new work disability instruments. *J Rehabil Med*. 2015 May;47(5):394-402. doi: 10.2340/16501977-1947.
8. Porcino J, Marfeo EE, McDonough CM, Chan L. The Work Disability Functional Assessment Battery (WD-FAB): Development and validation review. *TBV*. 2018, 26, 344.
9. NIH Clinical Center Rehabilitation Medicine Department. *Pilot Study to Assess the Use of the Work Disability Functional Assessment Battery (WD-FAB) in the Continuing Disability Review Business Process: Study Design Final Report to SSA*. October 2019.
10. Henly M, McDonough CM, Porcino J, Peterik K, Rasch EK, Marfeo EE, Houtenville AJ, Brucker DL. Linking job duties, functioning, and employment status using the Work-Disability Functional Assessment Battery (WD-FAB): An expert coding and quantitative analysis. *Work*. 2022 Sep 9.
11. NIH Clinical Center Rehabilitation Medicine Department. *Quarterly Report to the SSA 2023 Quarter 4: July 1, 2023 – September 30, 2023*. October 2023.
12. Ciol MA, Rasch EK, Hoffman JM, Huynh M, Chan L. Transitions in mobility, ADLs, and IADLs among working-age Medicare beneficiaries. *Disabil Health J*. 2014 Apr;7(2):206-15. doi: 10.1016/j.dhjo.2013.10.007.

Appendix A

Supplementary Tables for Describing Analytic Sample

Table A1. SSA Administrative Data for Analytic Sample vs. Non-participants

Sample Description	Analytic Sample	Non-respondent / Ineligible
Sample Count (n)	1244	14035
Percent of Participant Pool	7.56	85.27
Female (%)	**57.56**	46.54
Age in 2023 (mean)	**47.99**	44.89
Years on Benefits (mean)	3.96	3.93
MIE (%)	**27.97**	31.15
MIP (%)	59.32	57.68
MINE (%)	12.70	11.16
LOW (%)	20.90	16.46
MEDIUM (%)	22.51	19.33
HIGH (%)	**56.59**	64.21
SSI Only (%)	**60.13**	46.03
CPD Impairment – Physical (%)	**63.10**	56.01

Table A2. Demographics for Entire Analytic Sample and By Survey 2 Administration Mode

	Total		Telephone		Web	
	n / Mean	Percent	n / Mean	Percent	n / Mean	Percent
Study Sample	1244	100.0	309	24.8	935	75.2
Age at Survey 2	47.3	N/A	54.3	N/A	45.0	N/A
Gender						
Female	707	56.8	173	56.0	534	57.1
Male	531	42.7	136	44.0	395	42.2
Race & Ethnicity						
American Indian or Alaska Native	42	3.4	12	3.9	30	3.2
Asian	27	2.2	3	1.0	24	2.6
Black or African American	269	21.6	103	33.3	167	17.8
Native Hawaiian or Other Pacific Islander	10	0.8	2	0.7	8	0.9
White	812	65.2	172	55.7	640	68.4
Race: All other responses	31	2.5	0	0	31	3.3

Ethnicity: Hispanic	122	9.8	25	8.1	97	10.4
Marital Status						
Never Married	430	34.6	97	31.4	333	35.6
Married	306	24.6	70	22.6	236	25.2
Divorced	250	20.1	88	28.5	162	17.3
Education Level						
Less than high school	145	11.7	39	12.6	106	11.3
High school diploma or GED	441	35.4	116	37.5	325	34.7
More than high school	651	52.3	154	49.8	497	53.1
CPD Impairment						
Mental	459	36.9	89	28.8	370	39.6
Physical	785	63.1	220	71.2	565	60.4
CDR Impairment						
Mental	440	35.4	81	26.2	359	38.4
Physical	804	64.6	228	73.8	576	61.6

Appendix B

Supplementary Tables and Figures for Analyzing Single Administration WD-FAB Data

Table B1. Mean WD-FAB Scores by SSA Sub-Population (Survey 1)

SURVEY 1	BM	UBF	FMF	CM	CC	SR	RS	ME
TOTAL	48.6	46.1	62.4	45.6	47.6	51.3	45.3	50.1
TITLE II	48.7	46.2	62.3	45.9	47.7	51.7	45.5	50.4
TITLE XVI	48.0	45.5	62.3	44.7	47.4	50.7	45.0	49.7
CONCURRENT	53.9	50.0	66.2	47.4	47.7	49.3	45.0	48.2
CPD MENTAL	51.8	48.3	62.8	45.0	45.5	48.5	42.7	46.3
CPD PHYSICAL	46.8	44.8	62.2	45.9	48.8	52.9	46.9	52.3
CDR MENTAL	52.1	48.7	62.9	44.9	45.3	48.3	42.6	46.2
CDR PHYSICAL	46.7	44.7	62.2	46.0	48.9	52.9	46.8	52.2
CDR BS 1	44.1	42.6	61.1	44.6	48.1	51.8	46.1	50.7
CDR BS 11	47.1	44.3	60.2	45.6	47.4	52.8	46.0	51.8
CDR BS 13	48.2	46.0	62.5	46.6	49.2	54.1	48.5	53.9

Table B2. Mean WD-FAB Scores by SSA Sub-Population (Survey 2)

SURVEY 2	BM	UBF	FMF	CM	CC	SR	RS	ME
TOTAL	48.7	46.2	62.4	45.8	47.9	51.7	45.6	50.9
TITLE II	48.7	46.2	62.6	45.7	47.8	52.1	45.7	50.9
TITLE XVI	48.4	45.9	61.9	45.7	48.0	51.5	45.3	51.1
CONCURRENT	52.8	50.5	66.0	47.3	48.3	48.7	46.5	49.1
CPD MENTAL	51.9	48.7	62.8	45.2	45.9	48.9	43.1	47.1
CPD PHYSICAL	46.9	44.8	62.2	46.2	49.1	53.4	47.0	53.1
CDR MENTAL	52.3	49.0	62.8	45.1	45.8	48.7	42.9	46.9
CDR PHYSICAL	46.8	44.7	62.2	46.2	49.1	53.4	47.0	53.1
CDR BS 1	44.5	42.9	61.6	45.3	48.7	52.3	46.4	51.7
CDR BS 11	46.7	43.5	60.0	45.7	47.8	53.4	46.5	52.7
CDR BS 13	48.2	46.0	62.6	46.4	49.7	54.2	48.3	53.6

Table B3. Distribution of WD-FAB Profiles by SSA Sub-Population at Both Surveys

SURVEY 1	PROFILE 1 (%)	PROFILE 2 (%)	PROFILE 3 (%)	PROFILE 4 (%)
TOTAL	44.2	23.5	32.1	0.2
TITLE II	41.7	22.1	36.1	0.1
TITLE XVI	47.3	25.8	26.7	0.2
CONCURRENT	54.4	23.9	21.7	0.0
CPD MENTAL	62.3	21.8	15.7	0.2
CPD PHYSICAL	33.6	24.5	41.8	0.1
CDR MENTAL	64.3	20.2	15.2	0.2
CDR PHYSICAL	33.2	25.3	41.4	0.1
CDR BS 1	40.0	24.8	35.2	0.0
CDR BS 11	38.9	25.0	36.1	0.0
CDR BS 13	22.3	26.2	51.5	0.0

SURVEY 2				
TOTAL	40.4	25.8	33.7	0.0
TITLE II	38.8	25.8	35.4	0.0
TITLE XVI	42.4	25.6	32.0	0.0
CONCURRENT	47.8	28.3	23.9	0.0
CPD MENTAL	58.6	23.3	18.1	0.0
CPD PHYSICAL	29.8	27.3	42.9	0.0
CDR MENTAL	59.8	23.6	16.6	0.0
CDR PHYSICAL	29.9	27.0	43.2	0.0
CDR BS 1	35.6	28.9	35.6	0.0
CDR BS 11	38.0	23.1	38.9	0.0
CDR BS 13	20.0	20.8	59.2	0.0

Table B4. Mean WD-FAB Scale Scores and Distribution of Profiles by CDR Mailer Response Category (Survey 1)

SSA-455	WORK		HEALTH STATUS			HOSPITALIZATION		DR. VISIT	
RESPONSE	Yes	No	Better	Same	Worse	Yes	No	Yes	No
N	238	1006	117	550	575	651	592	1172	71
%	19.1	80.9	9.4	44.2	46.2	52.3	47.6	94.2	5.7
MEAN SCORE									
BM	52.4	47.7	53.4	50.6	45.7	47.9	49.5	48.3	53.5
UBF	49.6	45.3	50.1	48.0	43.5	45.5	46.8	45.9	50.1
FMF	63.7	62.2	65.6	63.7	60.6	62.1	62.8	62.4	63.4
CM	47.2	45.1	49.4	46.1	44.0	45.4	45.8	45.5	46.9
CC	48.7	47.4	51.0	48.2	46.3	47.6	47.7	47.5	49.0
SR	51.5	51.2	53.1	51.4	50.8	51.3	51.3	51.3	51.3
RS	47.0	44.9	48.3	45.7	44.3	45.4	45.3	45.3	46.3
ME	51.5	49.7	55.3	51.5	47.7	49.6	50.6	50.0	51.7
% PROFILE									
P1	15.8	84.2	5.3	40.2	54.4	53.5	46.5	94.0	5.8
P2	20.2	79.8	10.3	45.9	43.8	48.6	51.0	93.8	6.2
P3	22.8	77.2	14.5	48.3	37.0	53.5	46.5	94.8	5.2

Table B5. Mean WD-FAB Scale Scores and Distribution of Profiles by CDR Mailer Response Category (Survey 2)

SSA-455	WORK		HEALTH STATUS			HOSPITALIZATION		DR. VISIT	
RESPONSE	Yes	No	Better	Same	Worse	Yes	No	Yes	No
N	236	1008	126	556	560	616	626	1157	87
%	19.0	81.0	10.1	44.7	45.0	49.5	50.3	93.0	7.0
MEAN SCORE									
BM	52.6	47.8	54.8	50.2	45.9	47.3	50.1	48.3	54.0
UBF	49.9	45.4	51.2	47.8	43.6	45.1	47.4	45.9	50.4
FMF	64.1	62.1	66.4	63.2	60.8	62.0	62.9	62.3	64.2
CM	47.8	45.2	51.0	46.4	43.9	45.3	46.4	45.7	47.4
CC	49.0	47.7	50.8	48.3	46.9	47.7	48.2	47.8	49.1
SR	52.0	51.7	53.8	51.9	51.1	51.8	51.7	51.7	52.8
RS	46.9	45.2	49.0	45.8	44.5	45.6	45.5	45.5	46.6
ME	52.1	50.6	56.2	52.3	48.2	50.3	51.4	50.6	54.4

% PROFILE									
P1	15.7	84.3	4.0	40.0	56.0	51.5	48.3	93.6	6.4
P2	19.0	81.0	11.5	46.1	42.1	48.3	51.4	92.5	7.5
P3	22.9	77.1	16.4	49.3	34.1	48.1	51.9	92.6	7.4

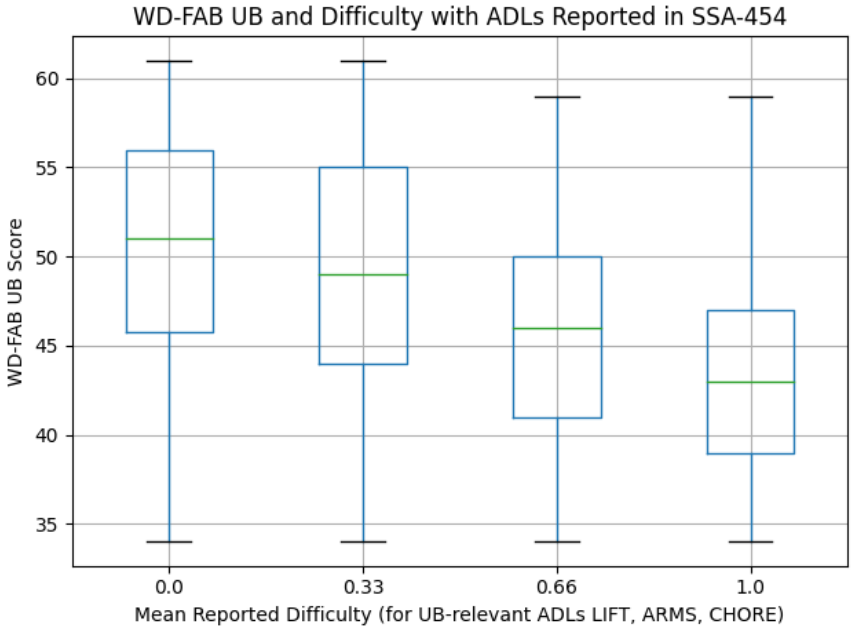

Figure B1. Box Plot of Upper Body Function scores by Mean Reported Difficulty with SSA-454 Daily Activities of Lifting Objects, Using Arms, and Doing Chores

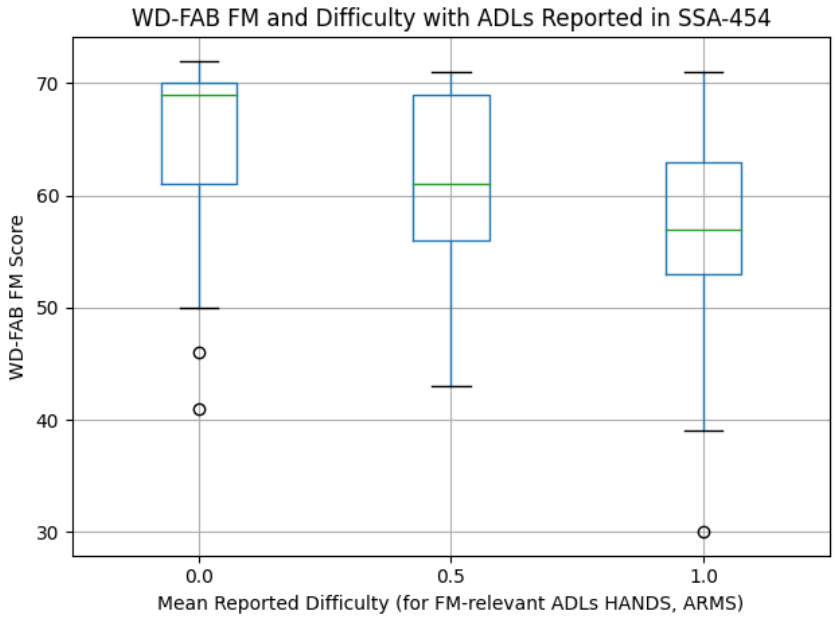

Figure B2. Box Plot of Fine Motor Function scores by Mean Reported Difficulty with SSA-454 Daily Activities of Using Hands or Fingers and Using Arms

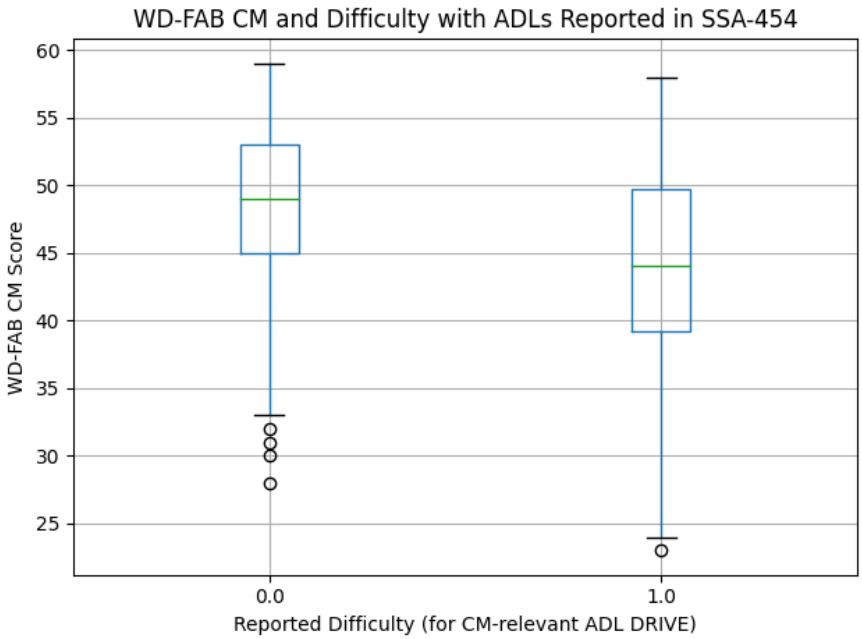

Figure B3. Box Plot of Community Mobility scores by Mean Reported Difficult with SSA-454 Daily Activity of Driving or Using Public Transportation

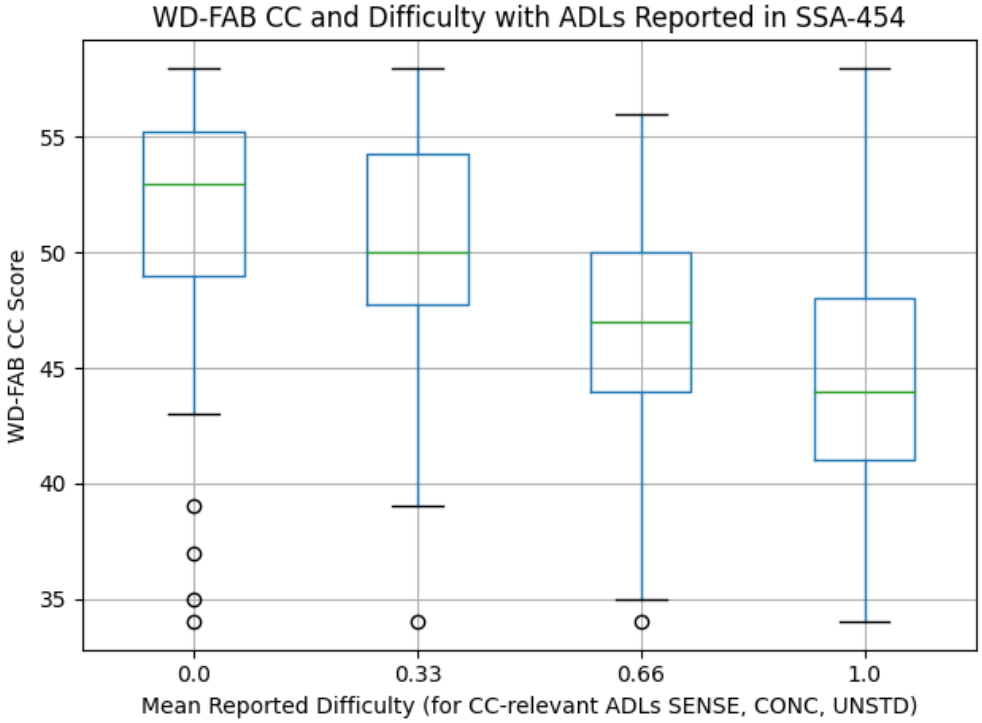

Figure B4. Box Plot of Communication & Cognition scores by Mean Reported Difficult with SSA-454 Daily Activities of Concentrating, Understanding or Following Directions, and Seeing, Hearing, or Speaking

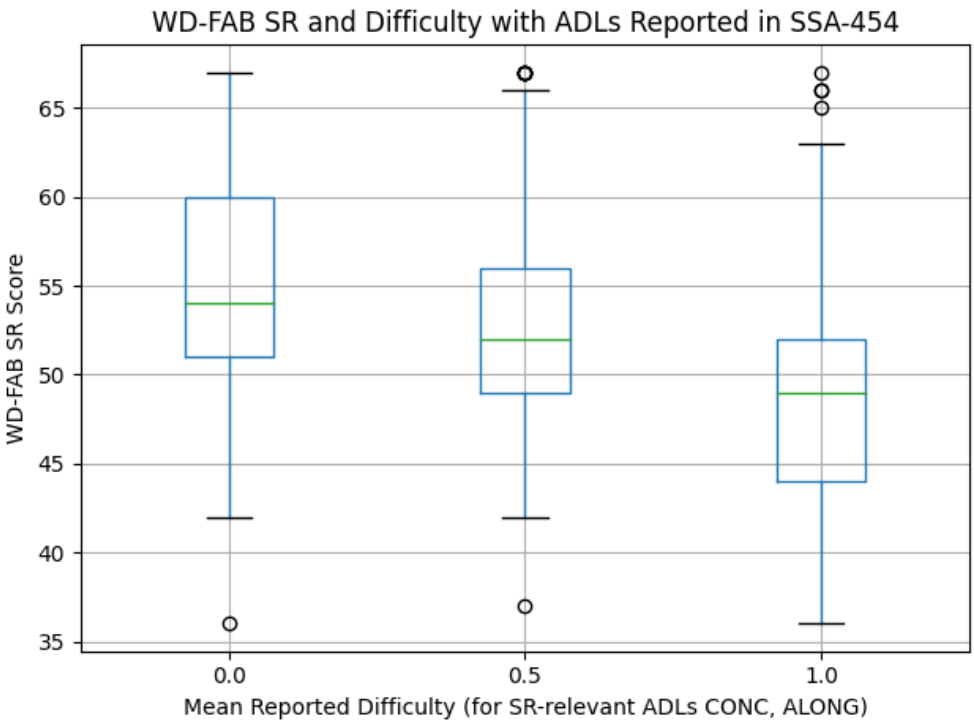

Figure B5. Box Plot of Self-Regulation scores by Mean Reported Difficult with SSA-454 Daily Activities of Concentrating and Getting Along with People

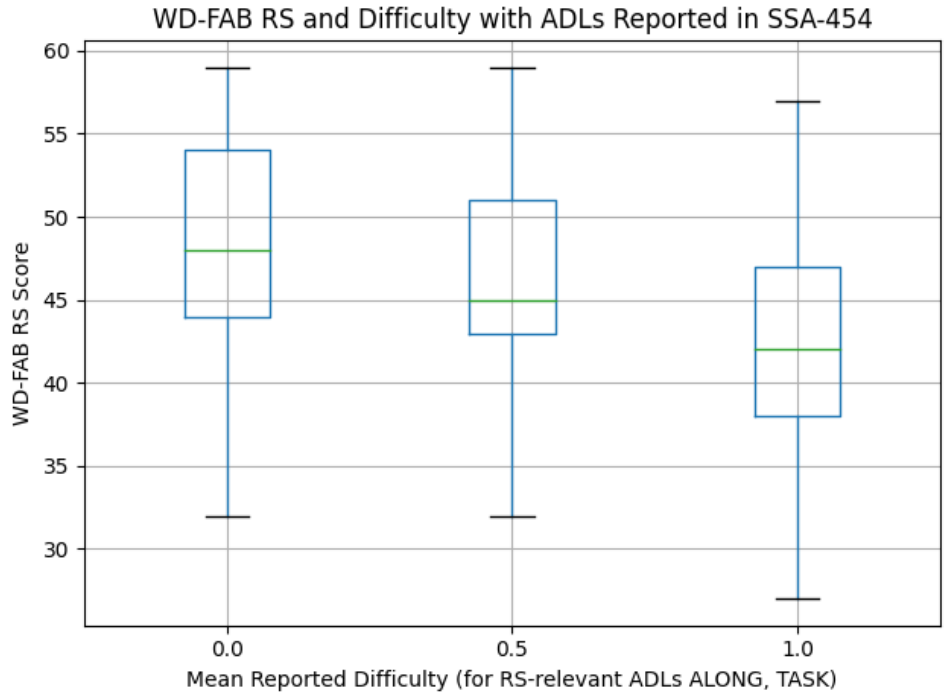

Figure B6. Box Plot of Resilience & Sociability scores by Mean Reported Difficult with SSA-454 Daily Activities of Getting Along with People and Completing Tasks

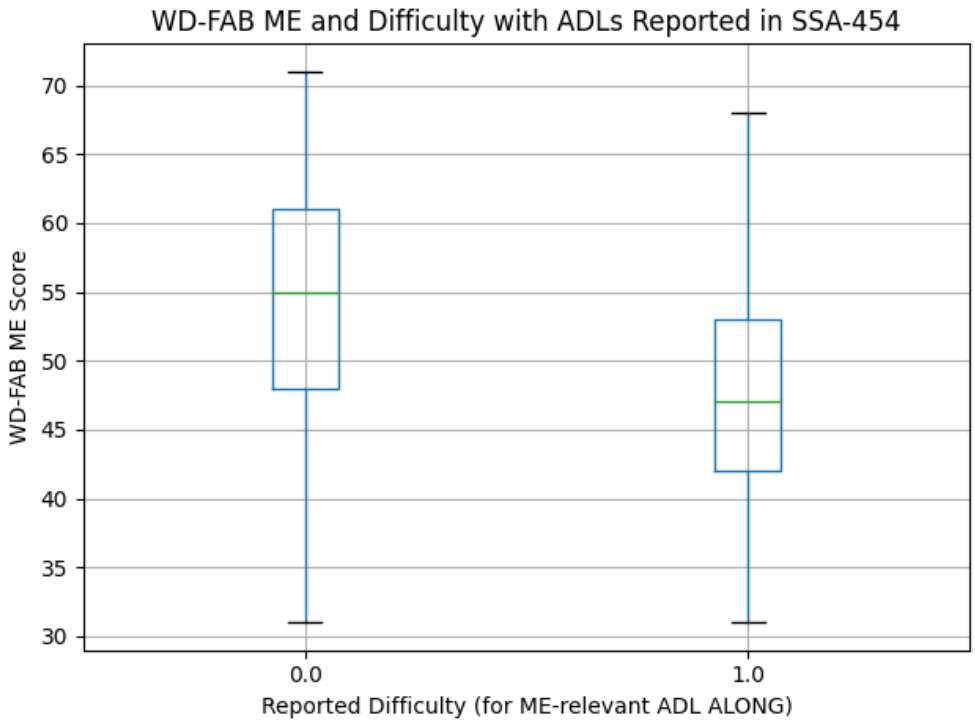

Figure B7. Box Plot of Mood & Emotion scores by Mean Reported Difficult with SSA-454 Daily Activity of Getting Along with People

Table B6. Mean WD-FAB Scores at Each Survey for Continuances vs. Cessations

	N	BM	UBF	FMF	CM	CC	SR	RS	ME
SURVEY 1									
CONTINUANCE	1162	48.4	45.9	62.2	45.5	47.6	51.2	45.3	50.0
CESSATION	82	51.5	49.1	65.9	46.9	48.3	51.9	46.1	51.5
SURVEY 2									
CONTINUANCE	1162	48.5	46.0	62.3	45.7	47.9	51.7	45.6	50.8
CESSATION	82	51.3	49.7	64.2	47.9	48.4	52.0	45.4	52.5

Appendix C

Bayesian Logistic Regression: Variables and Normalization Factors

Table C1. Variables Included in Bayesian Logistic Regression Modeling Results

Variable	Description
AGE	Age at time of CDR in years
F_LOBYRS_TOT	Total time on benefits in years
F_GENDER_FEMALE	Gender: Female
F_DIARYDESC_MIE	CDR Diary Type: MIE
F_DIARYDESC_MINE	CDR Diary Type: MINE
F_SSIORDI_SSDI	Program Type: SSDI
F_IMPTYPE_MENTAL	CPD Primary Impairment: Mental
S1_WDFAB_BM_NSC	Baseline (Survey 1) BM score
S1_WDFAB_UBF_NSC	Baseline (Survey 1) UBF score
S1_WDFAB_FMF_NSC	Baseline (Survey 1) FMF score
S1_WDFAB_CC_NSC	Baseline (Survey 1) CC score
S1_WDFAB_SR_NSC	Baseline (Survey 1) SR score
S1_WDFAB_RS_NSC	Baseline (Survey 1) RS score
S1_WDFAB_ME_NSC	Baseline (Survey 1) ME score
ΔBM	Change in BM score
ΔUBF	Change in UBF score
ΔFMF	Change in FMF score
ΔCC	Change in CC score
ΔSR	Change in SR score
ΔRS	Change in RS score
ΔME	Change in ME score
ΔPROFILE_DECLINE	Decline in WD-FAB Profile
ΔPROFILE_IMPROVE	Improvement in WD-FAB Profile
IMPROVE	Improvement in 1 or more WD-FAB scores (No Decline)
DECLINE	Decline in 1 or more WD-FAB scores (No Improvement)
MIXED	Both improvement and decline in WD-FAB scores

In all the Bayesian logistic regression models, we presented standardized effects. Table C2 presents normalization factors that one can use to interpret these effects

Table C2. Normalization Factors for Interpreting Standardized Effects

Variable	Description	Standard Deviation
F_LOBYRS_TOT	Total years on benefits	1.9 years
AGE	Age at CDR in years	11.7 years
S1_WDFAB_BM_NSC	Baseline BM score	8.2
S1_WDFAB_UBF_NSC	Baseline UBF score	7.2
S1_WDFAB_FMF_NSC	Baseline FMF score	7.5
S1_WDFAB_CC_NSC	Baseline CC score	6.1
S1_WDFAB_SR_NSC	Baseline SR score	6.8
S1_WDFAB_RS_NSC	Baseline RS score	7.1
S1_WDFAB_ME_NSC	Baseline ME score	9.6
ΔBM	Change in BM score	4.9
ΔUBF	Change in UBF score	4.8
ΔFMF	Change in FMF score	5.9
ΔCC	Change in CC score	4.2
ΔSR	Change in SR score	5.3
ΔRS	Change in RS score	5.4
ΔME	Change in ME score	6.5

If present in this table, one may convert the standardized effect for a given variable to the original units by dividing the effect by the given standard deviation. Any variables not found in the proceeding table are binary predictors for which we applied no transformations, and the standardized effect should be taken as-is. For example, the log-odds of having a decision code of CE increases on average by 0.19 per change in standardized Fine Motor Function (FMF) score. Using the normalization table, one can perform the conversion to find that the log-odds of cessation increases by 0.19/7.5 = 0.025 per point change of FMF on average.

Appendix D

Descriptive statistics for WD-FAB scale level administration data

Table D1. WD-FAB Administration Data for Physical Functioning Scales for Both Surveys

	Basic Mobility			Upper Body Function			Fine Motor Function			Community Mobility		
	Mean Time (min)	Median Time (min)	Mean # Items	Mean Time (min)	Median Time (min)	Mean # Items	Mean Time (min)	Median Time (min)	Mean # Items	Mean Time (min)	Median Time (min)	Mean # Items
Survey 1												
Total	1.58	1.43	6.99	1.75	1.58	8.11	1.26	1.12	7.97	1.88	1.78	9.03
Phone	1.99	1.93	6.91	2.15	2.04	7.82	1.63	1.55	7.93	2.47	2.39	8.92
Web	1.41	1.23	7.03	1.59	1.38	8.23	1.12	0.93	7.98	1.62	1.45	9.07
Survey 2												
Total	1.52	1.37	7.04	1.65	1.45	8.02	1.19	1.08	8.01	1.72	1.60	9.01
Phone	2.05	1.98	6.89	2.17	2.05	7.60	1.59	1.52	7.73	2.31	2.20	8.81
Web	1.35	1.18	7.09	1.50	1.27	8.15	1.07	0.92	8.10	1.52	1.33	9.08

Table D2. WD-FAB Administration Data for Mental Functioning Scales for Both Surveys

	Communication & Cognition			Self-Regulation			Resilience & Sociability			Mood & Emotions		
	Mean Time (min)	Median Time (min)	Mean # Items	Mean Time (min)	Median Time (min)	Mean # Items	Mean Time (min)	Median Time (min)	Mean # Items	Mean Time (min)	Median Time (min)	Mean # Items
Survey 1												
Total	2.17	2.03	10.19	1.93	1.80	9.30	2.02	1.83	10.61	1.55	1.42	8.65
Phone	2.77	2.68	9.91	2.36	2.23	9.06	2.27	2.13	9.82	1.98	1.90	8.64
Web	1.92	1.68	10.30	1.75	1.55	9.40	1.91	1.67	10.93	1.38	1.20	8.65
Survey 2												
Total	2.01	1.85	10.22	1.81	1.67	9.22	1.95	1.75	10.61	1.46	1.28	8.63
Phone	2.70	2.57	9.81	2.30	2.22	8.87	2.30	2.20	9.76	1.94	1.88	8.44
Web	1.79	1.57	10.35	1.66	1.43	9.34	1.83	1.58	10.89	1.31	1.10	8.68