Evaluation Tools for Human-AI Interactions Involving Older Adults with Mild Cognitive Impairments

Daisy M. Kiyemba
Jasmin Marwad
New England Robotics Validation and Experimentation (NERVE) Center,
University of Massachusetts Lowell
Lowell, Massachusetts, USA
daisy_kiyemba@student.uml.edu
jasmin_marwad@student.uml.edu

Elizabeth J. Carter
Robotics Institute, Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
ejcarter@andrew.cmu.edu

Adam Norton
New England Robotics Validation and Experimentation (NERVE) Center,
University of Massachusetts Lowell
Lowell, Massachusetts, USA
adam_norton@uml.edu

ABSTRACT
As artificial intelligence (AI) systems have already proven useful in human lives generally, there is an opportunity for specialized human-AI interaction (HAI) systems to support and provide care for older adults with mild cognitive impairment (MCI). However, the integration of this technology in this population must be thoughtfully designed to accommodate specific needs and limitations. This includes careful measurement of both humans and systems. We developed an evolving dataset categorizing relevant measurement tools into five groups: cognitive ability, demographics & personality, activity level, state of mind, and perceptions of the AI system. Each instance of the tool being used in the literature cataloged in the dataset is qualified in terms of how likely we would recommend using it in the domain of HAI for older adults with MCI based on contextual factors and internal reliability measures. This dataset will serve as a valuable resource for future research, aiding in the identification of promising areas and trends in AI systems for older adults with MCI as well as providing essential tools for future studies.

CCS CONCEPTS
• Computer systems organization → Robotics; • Human-centered computing → HCI design and evaluation methods.

KEYWORDS
human-AI interaction, mild cognitive impairment, evaluation

1 INTRODUCTION
Artificial intelligence (AI) systems—which include smart devices, wearable technology, and robots—are making significant strides in improving the quality of human life across many domains, including healthcare. Currently, AI is used for disease prediction, image analysis, drug discovery, and other applications (see [28, 29] for reviews of AI in healthcare). There are also efforts to incorporate AI into healthcare settings to support care provisions for older adults with mild cognitive impairment (MCI) and dementia. For example, one study examined the potential utilization of social assistive robots to aid therapists in training programs aimed at enhancing the cognitive function of older adults with MCI and mild dementia and found promising results [27]. Additional investigations have shown that engaging in social interactions with robots enhanced cognitive function in MCI patients [22]. Furthermore, advancements in technological applications, exemplified by the Brain m-App [21], have demonstrated notable enhancements in executive function and memory performance for people with cognitive frailty.

MCI impacts abilities such as memory, thinking, and problem-solving skills, while dementia causes more pronounced cognitive deficits including language, judgment, and reasoning. Progression of these conditions can significantly hinder a person’s ability to perform routine activities, such as remembering important appointments, communicating effectively, independently carrying out daily tasks, and maintaining personal hygiene. The application of AI in this domain is particularly prescient due to a growing older population and a projected diminishing population of young adults [2] to act as their informal caregivers (e.g., family members or friends), putting strain on existing caregiving services.

This situation presents an opportunity for human-AI interaction (HAI) systems to support care provision of older adults with MCI, including assistive technologies that have the capability to help with day-to-day activities, monitor health status, and provide remote assistance. However, HAI systems in this domain must be thoughtfully designed to accommodate the specific needs and limitations of individuals with MCI. Performance evaluations conducted for these technologies must be done using a set of qualified tools and validated measures, which can leverage those used in related fields such as medicine, psychology, computer science, robotics, and human factors. Various metrics and evaluation methods have been employed to assess people’s cognitive abilities, state of mind, and
subject perceptions of an AI agent when interacting with AI systems. In order to distill the many available tools and measures into a set that is qualified and validated, we have developed a dataset that collects the many types of assets available in the literature, categorized them based on a set of relevant characteristics, and qualified them according to a set of criteria. This paper presents the dataset which is intended to serve as a starting point when planning which tools to use during a study of HAI involving older adults with MCI. The paper is organized into five groups of tools that characterize: (1) cognitive ability; (2) demographics, personality, and experiences; (3) activity level; (4) state of mind; and (5) perceptions of the AI system.

2 METHODS

This dataset, while developed by reviewing relevant research literature, was not derived through a typical literature review or survey process, but rather followed a more flexible research method. To collect resources for the dataset, we searched numerous databases (PsycINFO, Google Scholar, IEEE Xplore, ACM, MEDLINE, PubMed, etc., using search terms including mild cognitive impairment or MCI, older adults, geriatric, ADLs, IADLs, EADLs, technology, AI systems, etc.) to identify studies and review types of publications in journals and conferences between the dates of 2000 to 2023. For the papers that contained extensive reviews of literature or cited original tools, we would further look into the citations of those papers. This process did take us beyond our limited data range; for example, the dataset has tools such as the Physical Self-Maintenance Scale (PSMS) and Instrumental Activities of Daily Living (IADL) Scale from a paper written in 1969 [13]. To date, we have identified 355 papers across the domains of medicine, psychology, computer science, robotics, and human factors, among others.

Next, members of the research team read each paper to identify the following: (1) whether it included a study; (2) whether and what type of technology was used; (3) whether participants included older adults and/or participants with a diagnosis of MCI (or other cognitive impairments such as dementia, Alzheimer’s, etc.). We recorded which individual evaluation tools were used by the research team; this downselection process resulted in 207 papers for which we characterized the tools used in those studies to form the dataset. We classified each citation that used each tool as to whether or not the documented study: (1) involved people interacting with AI/technology; (2) included older adults; (3) included participants with MCI or dementia; and (4) used the tool in its original or unmodified form. The tools used were then categorized into five groups to broadly distinguish their usage in a study: measuring (1) cognitive ability; (2) demographics, personality, and experiences; (3) activity level; (4) current state of mind; and (5) perceptions of the AI system.

Subsequently, we conducted an examination of their Cronbach’s α scores to assess internal reliability. We created tiers based on how likely we would be to recommend using each tool in the domain of HAI with older adults with MCI. Tier 1 included tools with Cronbach’s α ≥ 0.7 when used with older adults with MCI in experimental settings with or without AI interaction. Tools in this tier may require augmentation in order to be deployed successfully with MCI populations or when experimenting with AI systems, but minimally fit the criteria for use with older adults. As an exception, we designated certain tools as Tier * . These tools satisfy the criteria for Tier 1 (i.e.,hey have been utilized in experimental settings involving interactions between AI and older adults with MCI) but, to the best of our knowledge, lack reported Cronbach’s α scores. For certain tools, some studies may have employed and reported alternative reliability measures that were not included in our dataset. Lastly, all remaining tools that do not meet the criteria for Tier 1, 2, or * were assigned to Tier 3. It should be noted that a tool may be found in one or more tiers because multiple studies used the same tool yet resulted in varying reliability scores, contexts, etc. For example, the Geriatric Depression Scale (GDS) used in [19] was qualified as Tier 1 because it has a Cronbach’s α of 0.92 and has been used with older adults with MCI interacting with AI systems, whereas the same tool has also been used in [3] with a reported Cronbach’s α of 0.60 and for this reason, was placed in Tier 3.

Our methodology has some limitations. First, we opted to use Cronbach’s α as our primary reliability tool measure, considering its widespread use and straightforward nature compared to other types of reliability and validation methods. Second, due to the incorporation of external recommendations and the flexibility of our research approach, it is unlikely that the resulting dataset could be easily reproduced exactly. The dataset is intended to be a continuously updated resource as more tools are found and more relevant research is published. Readers are encouraged to contact the authors with recommended additions to the dataset (see the Zenodo page linked in Section 3 for contact information). To gain a more comprehensive understanding of how these tools have been applied in various contexts, we strongly encourage readers to consult the source papers/studies that are categorized in the dataset for more detail regarding utilization of the tools.

3 DATASET

The dataset can be accessed via Zenodo here: https://doi.org/10.5281/zenodo.8428760. Note that Zenodo includes a version history; please ensure you are accessing the latest version as the database will be continually updated. See Table 1 for statistics about the number of unique tools contained in the dataset in each category and tier; note that these are accurate as of January 2024, the final submission date of this paper. Otherwise, for most up-to-date results, see the current Zenodo dataset at the link above. A short summary of each category can be found below.

Characterizing Cognitive Ability. In the dataset, we defined the cognitive ability category of tools as those used to assess anything related to cognition such as performance, decline or impairment, memory, etc. Tools in this category include Mini Mental State Examination (MMSE) [19], Montreal Cognitive Assessment (MoCA) [18], Clinical Dementia Rating (CDR), Direct Assessment of Functional Status (DAFS) [12] and other neuropsychological batteries such as Frontal Assessment Battery (FAB) [16].
Table 1: Total number of tools in each category and tier. Note: The total number of unique tools across tiers varies from the sum of the individual values per category-tier combination due to some tools with multiple entries in the dataset (i.e., multiple examples of the studies that used the same tool) wherein they were used in contexts that each qualified as different tiers.

<table>
<thead>
<tr>
<th>Category</th>
<th>Total Unique Tools (N=247)</th>
<th>Tier 1</th>
<th>Tier 2</th>
<th>Tier 3</th>
<th>Tier *</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive Ability</td>
<td>80</td>
<td>6</td>
<td>20</td>
<td>36</td>
<td>45</td>
</tr>
<tr>
<td>Demographics, Personality, and Experiences</td>
<td>10</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>Activity Level</td>
<td>31</td>
<td>5</td>
<td>11</td>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td>State of Mind</td>
<td>37</td>
<td>2</td>
<td>15</td>
<td>16</td>
<td>18</td>
</tr>
<tr>
<td>Perceptions of the AI</td>
<td>89</td>
<td>15</td>
<td>33</td>
<td>21</td>
<td>49</td>
</tr>
<tr>
<td>Total</td>
<td>28</td>
<td>82</td>
<td>86</td>
<td>131</td>
<td></td>
</tr>
</tbody>
</table>

Characterizing Demographics, Personality, and Experiences. Under this category, we classified tools used to assess demographics, personality traits, and/or general experiences. Examples of tools in this category include the McGill Friendship Questionnaire [10], The Revised NEO Personality Inventory (NEO-PI-R) [4], and World Health Organisation Five Well-Being Index (WHO-5) [17].

Characterizing Activity Level. This category includes tools designed to assess a person’s activity level characterized by an individual’s level of independence, level of difficulty in performing tasks, and the degree of assistance they require. Some of the tools include Everyday Compensation (EComp) Questionnaire [26], Functional Activities Questionnaire (FAQ) [8], Medication Management Capacity (MMC) [24], Performance-Based Skills Assessment Financial Skills subscale (UPSA Finances) [20], and Caregiver Assisting ADL Scale [5].

Characterizing State of Mind. Tools used to assess mood, emotion, and mental state were classified in the category of state of mind. Some examples include Geriatric Depression Scale (GDS) [3], Caregiver Burden Scale (CBS) [6], Spielberger state-trait anxiety inventory (STAI) [7], Apathy Evaluation Scale (AESD) [11], and Positive and Negative Affect Schedule (PANAS) [14].

Assessing Perceptions of the AI. In this last category, we included tools used to assess perceptions of AI systems during or after an interaction and/or general attitudes towards technology, for example, Robot Attitudes Scale (RAS) [25], Godspeed questionnaire [1], Unified Theory of Acceptance and Use of Technology (UTAUT) [9], Trust in Technology [15], and Negative Attitudes towards Robots Scale (NARS) [23].

4 USAGE NOTES

While we have included numerous tools in our dataset, it is imperative to acknowledge that several of them, particularly those in Tier 2, 3, and * require further research and analysis before they can be recommended for usage in the target domain. These tools have yet to undergo validation through experimentation with AI systems and/or the target population of older adults with MCI, and/or they lack a reported Cronbach’s α score. Readers should also be mindful of variations in reported Cronbach’s α scores for certain tools, which can result from modifications made in specific studies, when adapted to different languages, or used in different contexts. Furthermore, the effective use of specific tools, particularly those related to neuropsychological assessment, require training or, at the very least, a level of familiarity and experience in their application. A column in the dataset is included to highlight the tools that may require training or more substantial level of familiarity in the related field before the tool can be effectively used. The dataset is intended to provide a quick reference when planning a study of HAI involving older adults with MCI, but further detail regarding the utilization of each tool is required before doing so.

Additionally, the context under which each tool was developed, validated, and used in each particular study must be considered when making decisions to adopt it for another study (i.e., factors like language, culture, demographics, etc., will impact the successful usage of the tool). Relevant examples from the dataset include the Hong Kong version of the MoCa (Cronbach’s α of 0.60 [3]) compared to another instance that is 0.83 [18]) and the Korean version of the GDS (Cronbach’s α of 0.92 [19] compared to another instance that is 0.60 [3]). These tools have been translated into other languages differing from that of their original developments illustrating how language is a factor that can affect reliability scores.

5 FUTURE WORK

We aim to compose a comprehensive follow-up analytical paper in the near future. Furthermore, it is important to note that the development of this dataset is anticipated to undergo continuous evolution and refinement over time whereby we will continually update the dataset as needed. Readers are encouraged to contact the authors of this paper and dataset (using the information on the Zenodo page) to recommend additional tools and entries to the dataset. We also recommend additional research to be conducted to validate the tools using the population, especially the tools in lower tiers that have a low Cronbach’s α and haven’t been used with this specific population or context. This will provide valuable insights into the evolving landscape of these tools and their potential contributions to the field.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (IIS-2112633).

REFERENCES


