Calibration of the Activity Inventory Item Bank: A Patient-Reported Outcome Measurement Instrument for Low Vision Rehabilitation

Micaela Gobeille¹, Chris Bradley¹, Judith E. Goldstein¹, and Robert Massof¹

¹ Johns Hopkins University School of Medicine, Baltimore, MD, USA

Purpose: To provide calibrated item measures and rating category thresholds for the Activity Inventory (AI), an adaptive visual function questionnaire, from difficulty ratings obtained from a large sample of new low vision patients at pre-rehabilitation baseline.

Methods: Baseline AI (510 items) rating scale data from five previous low vision rehabilitation outcome studies (n = 3623) were combined, and the method of successive dichotomizations was used to estimate calibrated item measures and rating category thresholds. Infit statistics were analyzed to evaluate the fit of the data to the model. Factor analysis was applied to person measures estimated from different subsets of items (e.g., functional domains such as reading, mobility) to evaluate differential person functioning.

Results: Estimated item measures were well targeted to the low vision patient population. The distribution of infit statistics confirmed the validity of the estimated measures and the two-factor structure previously observed for the AI.

Conclusions: Our calibrated item measures and rating category thresholds enable researchers to estimate changes in visual ability from low vision rehabilitation on the same scale, facilitating comparisons between studies.

Translational Relevance: The work described in this paper provides calibrated item measures and rating category thresholds for a visual function questionnaire to measure patient-centered outcomes in low vision clinical research. The calibrated AI also can be used as a patient outcome measure and quality assurance tool in clinical practice.

Introduction

Over the past few decades, there has been a proliferation of rating scale questionnaires to evaluate self-reported vision-related functional outcomes. The Activity Inventory (AI) is one of the few visual function questionnaires developed specifically for low vision rehabilitation (LVR), providing a patient-centered approach to LVR outcomes measurement.¹-⁴ The current version of the AI has an item bank of 510 items organized in a hierarchy of 50 goals (e.g., using a public restroom without assistance, preparing daily meals) and 460 underlying tasks (e.g., finding food items, avoiding burning oneself). The full AI provides a means to estimate visual ability, and subsets of related tasks within the AI (i.e., reading, mobility, visual information, visual motor) can be examined to assess visual ability within specific functional domains. The full AI presents different subsets of items according to an algorithm based on patient-specific preferences or responses. To compare estimated measures of patient ability from different studies on the same scale, this adaptive technique requires item calibrations that were last made available in 2005.⁴ In this paper, we provide item calibrations based on improved estimation methods, for the most recent AI item bank, from a large sample of patients with low vision.

To estimate visual ability and the effects of rehabilitation on an interval scale, Rasch analysis must be employed.¹ With Rasch analysis, three types of parameters are estimated: person measures, item measures, and rating category thresholds. Person measures
represent the amount of visual ability possessed by persons (in our case, the patients’ ability to perform vision-dependent activities), and item measures represent the amount of visual ability required by the items (i.e., the activities). More positive person measures reflect greater visual ability, and more positive item measures reflect greater visual ability required to perform the activity described by the item. When completing the AI or any visual function questionnaire, persons compare their ability (person measure) to the ability required to accomplish the item (item measure), thereby judging the magnitude of functional reserve—the difference between the person measure and item measure. Each functional reserve estimate for a person–item combination falls on the real number line in units of visual ability. Each point on the real number line belongs to a discrete interval, or response category, which is separated from adjacent intervals by rating category thresholds.

Previous studies with the AI estimated person measures, item measures, and rating category thresholds using the Andrich Rating Scale Model, which is a popular polytomous Rasch model. However, a mathematical property of the Andrich model (its multiplicative structure) leads to the frequent estimation of disordered rating category thresholds, which means that it in principle does not apply to rating scale data where thresholds must be ordered. Users of the Andrich model have been advised to merge neighboring rating categories when necessary, reducing the number of thresholds in the process, until all remaining estimated thresholds are ordered. Merging rating categories leads to a rating scale with fewer rating categories than used by the sample of persons. Additionally, merging rating categories in the Andrich model results in the unit of measurement changing (i.e., a scale change), which precludes direct comparison of estimated measures between studies or between baseline and follow-up within the same study. The method of successive dichotomizations (MSD) corrects this problem by directly extending the dichotomous Rasch model to multiple rating categories and estimates ordered thresholds on the same scale irrespective of the number of response categories. MSD has been shown through simulations to estimate person measures, item measures, and response category thresholds that are in near perfect agreement with the true values.

Although some interventions may improve the person (e.g., by improving vision through refractive error correction, cataract extraction), the majority of LVR interventions change the item by making it easier to perform the activity or by reducing the need to perform it (e.g., buying pre-cut vegetables rather than cutting them). These interventions may target selected activities, such as reading with magnification, using a white cane for mobility, or installing tactile markings on home appliances to facilitate cooking. Item-specific intervention for a given person differs from intervention that changes the person; the former results in different changes in the person’s functional reserve for different items, whereas the latter causes an identical magnitude change in the person’s functional reserve for all items. Ideally, a method of analysis could tease apart the effects of these two types of interventions. However, currently used conjoint estimation methods in Rasch analysis are limited to estimating a single person measure for all items and a single item measure for all persons. As a result, it is necessary to employ item anchoring, which uses item measures calibrated from the responses of a large representative sample, forcing all change in functional reserve into the person measure.

The goal of this paper is to provide a set of calibrated item measures and rating category thresholds for the AI. Clinicians and researchers can use these calibrations as anchors to evaluate visual ability outcomes for persons with low vision using the current version of the AI. MSD is applied to a large and representative sample of low vision outpatients with data pooled from five low vision outcome studies.

Methods

Database Construction

Baseline data were collected from five previous studies and included new patients with low vision (i.e., no low vision examination in prior 3 years) across the United States. All studies were approved by their respective institutional review boards and conformed to the tenets of the Declaration of Helsinki. Each study obtained informed consent from participating persons. These studies (Table) examined the community-dwelling low vision population, including a private outpatient clinic low vision population, low vision outpatients at high risk for depression who were recruited to participate in a depression-prevention clinical trial, home healthcare patients with a secondary diagnosis of low vision, low vision outpatients receiving rehabilitation services in a mobile clinic, and low vision outpatients recruited to participate in a clinical outcome study evaluating electronic head-mounted vision enhancement systems. AI data in these studies were collected by telephone interview prior to initial evaluation as detailed in their respective study protocols.

Baseline data from these studies
were pooled to create a large \((n = 3623)\) database to represent the community-dwelling low vision outpatient population.

**Activity Inventory**

The AI consists of 510 items (i.e., 50 goals and 460 underlying tasks) that persons respond to using six ordered response categories: not difficult, slightly difficult, moderately difficult, very difficult, extremely difficult but possible, and impossible to do without help. The AI employs an adaptive algorithm that determines which goals and tasks are both important and difficult for the patient.2–4 First the person is asked to rate the importance of a goal. Only if the goal exceeds some criterion level of importance does the person then rate the difficulty of the goal. If the person’s difficulty rating of the goal exceeds some criterion level, then the person is required to rate the difficulties of the underlying tasks of the goal or identifies those tasks as not applicable. Goals that are rated not important and, by extension, their subsidiary tasks are not included in the patient’s individualized rehabilitation plan and are therefore excluded from the analysis; this is also true of tasks that are rated not applicable regardless of the importance of the goal. Additionally, items reported as “not difficult” at baseline are filtered out in the analysis, as these items are not relevant to the rehabilitation plan, a process referred to as “item filtering.”7

Difficulty ratings are assigned an ordinal score before undergoing Rasch analysis, and the order of these scores determines whether an ability or disability scale is used. Our calibrations were estimated on an ability scale, with “impossible to do without help” items assigned a score of 0, “extremely difficult but possible” assigned a score of 1, “very difficult” assigned a score of 2, “moderately difficult” assigned a score of 3, “slightly difficult” assigned a score of 4, and “not difficult” assigned a score of 5.

### Table. Person Characteristics

| Dataset     | Sample Size, \(n\) | Female, % | Visual Acuity (logMAR) |
|-------------|---------------------|-----------|------------------------|
| Legacy8,9   | 2398                | 58.0      | 0.55                   |
| LVROS9      | 779                 | 67.0      | 0.6                    |
| VITAL10,11  | 212                 | 70.2      | 0.61                   |
| CLVOM13     | 158                 | 67.5      | 0.7                    |
| eLVES14     | 32                  | 40.6      | 1.0                    |
| MCOS12      | 70                  | 64.0      | 0.9                    |

Median Interquartile Range Median Age (y)

- Legacy: 0.3–0.9, Median Age: 71.6
- LVROS: 0.40–0.90, Median Age: 76.4
- VITAL: 0.41–0.78, Median Age: 84.5
- CLVOM: 0.40–1.0, Median Age: 79.0
- eLVES: 0.79–1.1, Median Age: 55.0
- MCOS: 1.1–0.60, Median Age: 59.5

**Analysis**

The R package msd (R Foundation for Statistical Computing, Vienna, Austria) was used to estimate item measures, rating category thresholds, and person measures, given the ratings of the AI items assigned by the 3623 patients. Two functions in the R msd package were used: *msd* and *pms*. The function *msd* was used to estimate calibrated item measures and rating category thresholds and to calculate their standard errors. By convention, the origin of the scale was defined to be the mean of the estimated item measures. Calibrating item measures and rating category thresholds allows researchers to estimate person measures from different samples of patients with low vision on the same scale. The function *pms* was then used to estimate person measures (and their standard errors) given the estimated calibrated item measures and rating category thresholds from the previous step (i.e., item measures and rating category thresholds were anchored). As explained in the introduction, anchoring item measures and thresholds forces all change in the functional reserve into the person measure. Person measures were estimated from difficulty ratings at both the goal and task levels, as well as for each functional domain (i.e., reading, mobility, visual information, and visual motor)2–4 Item measures, person measures, and rating category thresholds were also estimated using the Andrich Rating Scale Model in Winsteps® statistical software to compare the measures estimated by MSD to those estimated in the previous studies that supplied the item response data. Fit statistics were also estimated by both functions.

To determine the validity of the estimated measures it is necessary to assess the fit of the data to the Rasch model. This can be done using the infit mean square which, like the Rasch model, assumes that random variability is unidimensional and approximately normally distributed. A key assumption of the Rasch model is that the distribution of deviates
from the estimated measures is approximately normal, with the same variance for every person/item/threshold combination. To test this hypothesis, a mean squared fit statistic referred to as the infit mean square (information-weighted fit statistic) was used. The infit mean square statistic is expected to be distributed across persons and items as $\chi^2$ divided by the degrees of freedom, and it ranges from 0 to infinity with an expected value of 1. For any person or item, an infit mean square between 0 and 1 indicates that the error variance is less than expected by the model, whereas an infit mean square between 1 and infinity indicates that the error variance is greater than expected by the model. Because each patient completing the AI responds to a different number of items (i.e., different degrees of freedom for each person), we used a weighted sum $\chi^2$ model, each with its own degrees of freedom to evaluate the person infit mean square.

The infit mean square of item measure estimates can be influenced by differential person functioning (the person measure counterpart to differential item functioning), which occurs when the estimated person measures are different for different subsets of items. Consistent with previous work, we examined correlations of person measures estimated from responses to different subsets of items organized by the relevant functional domain (e.g., reading, mobility, visual information, visual motor). We also performed factor analysis on the covariance matrix for the different person measure estimates (from the four functional domains and goals) to look for evidence of multidimensionality in the estimate of the person measure.

**Results**

**AI Calibration**

Figure 1 shows a Wright map, a distribution of the calibrated item measures (gray bars below the line) compared to the estimated person measures (red bars above the line). Calibrated item measures ranged from −4.0 to 5.2 logits (SD = 1.60 logits) with a mean of 0 logit (by convention), and estimated person measures ranged from −4.1 to 6.2 logits (SD = 1.28 logits) with a mean of 0.7 logit. For a large portion of the range, item measures were well targeted to person measures, except at the tails of the distributions.

Figure 2 plots the precision of the estimates (i.e., the standard errors) against the corresponding estimated parameter values (i.e., the estimated person or item measures). The standard error of the person measure estimate is governed by the number of items a person responds to, the location of the person relative to the items, and the variability between people in their interpretation of the item content, which is independent of a difference in their abilities; a similar result is true for standard errors of the item measure estimate if one reverses the role of person and item. Greater error is expected for extreme item and person measures, resulting in the U-shaped distributions seen in Figure 2. The average standard error for the item measures was 0.027, or 0.017% of the variance of the items, which shows that the item calibrations are highly precise (due to the large number of persons in our sample).

Figure 3 shows a histogram of person infit mean squares (gray bars) and a weighted sum of $\chi^2$/degrees of freedom distributions (red line). More person infit mean squares fall to the right of 1 (i.e., actual variance is greater than expected) than to
Calibration of the Activity Inventory Item Bank

Figure 3. Person infit mean square and expected weighted sum of $\chi^2$ distributions. Person infit mean squares (gray bars) are plotted relative to a weighted sum of $\chi^2$ distributions (red line). Person infit mean squares more frequently fall in the right tail, with fewer infit mean squares falling around the expected value of 1.

Figure 4. Item infit z-score. (A) Item measures are plotted relative to item infit z-scores for each functional domain and for goals, with most z-scores falling beyond 2 SDs (indicated by black lines) from the expected mean and thus considered outliers. (B) The cumulative frequency is plotted. A cumulative frequency of 50% is achieved at an infit z-score of –3.06 in reading, 0.95 in visual information, 0.97 for goals, 4.67 for visual motor, and 6.84 for mobility. Thus, less variance than expected is observed for reading, variance is close to expected for visual information and goals, and more variance than expected is observed for visual motor and mobility.

Factor analysis was performed on person measures estimated from the difficulty ratings of tasks from each of the functional domains and of goals, in order to further investigate differences in item infit mean square $z$-score distributions. One factor was sufficient to explain 67.2% of all variance, whereas subsequent components contributed minimal amounts (12%, 9%, 6%, and 5% for the second through fifth components, respectively). However, two components, which explain 79.2% of all variance, are needed to explain the correlations between functional domains. The amount of variance explained in each domain was 68% in reading, 59% in goals, 74% in visual information, 60% in visual motor, and 67% in mobility. As seen in Figure 5, reading loaded more heavily on one factor whereas mobility loaded more heavily on the other, and remaining domains fell between these two.

Comparison of Estimation Methods

Figure 6 compares MSD and Andrich model parameters estimated from our data. Figure 6A compares the estimated item measures, which were highly correlated ($R = 0.98$) with systematic deviation only at the extremes. However, the slope of the regression line ($m = 0.36$) was far from 1, which reflects the scale differences between the two models—the Andrich model changes its scale as a function of the number of rating categories while MSD does not. Estimated rating category thresholds were also different, as seen in Figure 6B. Thresholds estimated from the Andrich
Calibration of the Activity Inventory Item Bank

**Figure 5.** Factor analysis for functional domains. Factor loadings are plotted for each functional domain. Reading loaded more heavily on factor 1, with 90% of explained variance attributed to factor 1 and 10% of explained variance attributed to factor 2. Goals loaded more heavily on factor 1, with 80% of explained variance attributed to factor 1 and 19% of explained variance attributed to factor 2. Visual information (73% of explained variance attributed to factor 1, 27% of explained variance attributed to factor 2) and visual motor (67% of explained variance attributed to factor 1, 33% of explained variance attributed to factor 2) fell between factor 1 and factor 2. Mobility loaded more heavily on factor 2, with 15% of explained variance attributed to factor 1 and 85% of explained variance attributed to factor 2.

The model were disordered (disordered thresholds do not define a rating scale), whereas thresholds estimated from MSD were ordered. **Figure 6C** compares person measures estimated by each model with item measures and rating category thresholds anchored. The correlation between estimated person measures is $R = 0.85$.

**Discussion**

This paper provides calibrated item measures and rating category thresholds for the existing AI, which can be used to estimate individualized baseline and post-intervention visual ability outcome measures for any sample of persons with vision impairment. To use the calibrated item measures and rating category thresholds, the `pms` function in the R package `msd` should be run with both the item measures and thresholds anchored (i.e., as inputs to the program). This way only the person measures are estimated. Ordinal AI data that are used to estimate person measures must be scored on an ability scale (i.e., “not difficult” is scored as 5, “impossible to do without help” is scored as 0). We provide instructions on how to use the `pms` function and also provide calibrated item measures and rating category thresholds stored as an R variable for convenience (https://sourceforge.net/projects/ai-calibrations/files/). Using these calibrated item measures and rating category thresholds allows visual ability estimates in different studies and/or different time points to be estimated on the same measurement scale. The updated calibrations based on MSD improve on previous work that used the Andrich Rating Scale Model, which often estimates disordered thresholds. With MSD it is not necessary to collapse rating categories, which reduces the resolution of outcome measures with the instrument. However, as shown in **Figure 6**, both models estimate person and item measures that are linearly related to each other. The present calibration is also based on a larger sample of 3623 patients, which improves resolution of item measure estimates for the 460 tasks that are administered adaptively.

Consistent with previous work, visual ability estimated from different subsets of items was observed to be bidimensional—two factors are necessary and sufficient to explain the covariances and differential person functioning in the data. These findings are consistent with previous studies suggesting that the two factors are explained by the independent effects of visual acuity and visual field loss/scotomas on visual ability. Visual acuity loss impacts activities related to identification and recognition (e.g., reading); visual field loss/scotomas have a greater impact on the visual perception of spatial relations (e.g., mobility function).

Our calibrated item bank enables subsets of AI items to be administered, either in an adaptive format or as a fixed item questionnaire; in both cases, the overall time of administration for an otherwise long questionnaire is reduced. For example, items from a single functional domain can be administered to target an intervention with a specified expected rehabilitation outcome (e.g., items from the reading domain could be used to assess outcomes involving magnification). However, if an intervention is expected to improve overall visual ability, the adaptive AI should be used.

Although the broad range of items composing the AI enable better precision in measuring LVR outcomes, the item bank still must be regularly updated to reflect changes in lifestyles, technology, and preferences of persons with low vision. For example, obsolete items should be eliminated and newly relevant items added (newly added items should nest under the appropriate objective and, if applicable, goal). To maintain a consistent scale, the AI should not be recalibrated when items are retired or added. To estimate item measures for the new items the following protocol should be applied:
(1) Administer the current AI (without new items) to a sample of people with low vision who have a distribution of traits similar to those in the Table. (2) Use the function pms in the R package msd to estimate person measures while anchoring item measures and rating category thresholds to the values in Supplementary Materials S1 and S2. (3) Have each subject rate the difficulty of each new item. (4) Use the function ins in the R package msd to estimate item measures for the new items by anchoring person measures to those estimated in step 2 and thresholds to the values in Supplementary Material S2. It is recommended that sample size be large enough so that the standard errors of the newly estimated item measures are comparable to those in Supplementary Material S1 (note that standard errors are inversely proportional to the square root of the number of responses to the item). The current version of the AI is well suited for assessing LVR outcomes; however, ongoing item bank modifications are needed to ensure its continued usefulness in measuring person-centered outcomes. We also note the limitation that our calibrations are based on a diverse low vision rehabilitation population and may not apply to all study populations.

This project calibrated the AI using MSD to provide item measures and rating category thresholds for evaluating low vision rehabilitation outcomes. Validity of the estimated measures was affirmed based on infit and standard error distributions. MSD and Andrich Rating Scale Model parameters (excluding rating category thresholds) were linearly related to each other. However, only MSD estimated ordered rating category thresholds and estimated all parameters on an invariant scale.

Acknowledgments

The authors thank the following for contributing data from previously published studies that were used in this project: Barry Rovner, MD; Robin Casten, PhD; and the Low Vision Research Network Study Group. We also thank Theresa Smith, PhD, OTR; Guy Davis; and members of the Comparison of Low Vision Outcome Measures study group: Lisa Foret, MS; Joan Gillard, MS; Kristen Shiflett, MS; Lind Stevens, MS; Rebecca O’Bryan, MS; Ben Hoagland, MS; Michelle Bianchi, MS; Robert Trahan, MS; Emily Murphy, MS; James Saba, MS; and Jean McBride, MS.

Supported by grants from the National Eye Institute, National Institutes of Health (R44EY028077, R01EY026617, R01EY022322, R34EY018696, R01EY012045, U01EY015839, and T35EY007149).

Disclosure: M. Gobeille, None; C. Bradley, None; J. Goldstein, None; R. Massof, None

References

1. Massof R. Understanding Rasch and item response theory models: applications to the estimation and validation of interval latent trait measures from responses to rating scale questionnaires. *Ophthalmic Epidemiol.* 2011;18(1):1–19.
2. Massof R, Ahmadian L, Grover L, et al. The Activity Inventory: an adaptive visual function questionnaire. *Optom Vis Sci.* 2007;84(8):763–774.
3. Massof R, Hsu C, Baker F, et al. Visual disability variables. I: The importance and difficulty of activity goals for a sample of low vision patients. *Arch Phys Med Rehabil.* 2005;86(5):946–953.
4. Massof R, Hsu C, Baker F, et al. Visual disability variables. II: The difficulty of tasks for a sample of low vision patients. *Arch Phys Med Rehabil.* 2005;86(5):954–967.
5. Bradley C, Massof R. Method of successive dichotomizations: an improved method for
estimating measures of latent variables from rating scale data. *PLoS One.* 2018;13(10):e0206106.

6. Bond TG, Fox CM. *Applying the Rasch model: Fundamental Measurement in the Human Sciences.* 3rd ed. London: Routledge; 2015.

7. Massof R, Stelmac J. Interpretation of low vision rehabilitation outcomes measures. *Optom Vis Sci.* 2013;90(8):788–798.

8. Goldstein J, Jackson M, Fox S, et al. Clinically meaningful rehabilitation outcomes of low vision patients served by outpatient clinical centers. *JAMA Ophthalmol.* 2015;133(7):E1–E8.

9. Goldstein J, Chun M, Fletcher D, et al. Visual ability of patients seeking outpatient low vision services in the United States. *JAMA Ophthalmol.* 2014;132(10):1169–1177.

10. Deemer A, Massof R, Rovner B, et al. Functional outcomes of the low vision depression prevention trial in age-related macular degeneration. *Invest Ophthalmol Vis Sci.* 2017;58(3):1514–1520.

11. Rovner B, Casten R, Hegel M, et al. Low vision depression prevention trial in age-related macular degeneration: a randomized clinical trial. *Ophthalmology.* 2014;121(11):2204–2211.

12. Gobeille M, Malkin A, Jamara R, et al. Clinical outcomes of low vision rehabilitation delivered by a mobile clinic. *Ophthalmic Physiol Opt.* 2018;38(2):193–202.

13. Massof R, Smith T, Malkin A, et al. Comparison of low vision rehabilitation outcome measures. *Invest Ophthalmol Vis Sci.* 2015;56(7):494.

14. Deemer A, Swenor B, Fujiwara K, et al. Preliminary evaluation of two digital image processing strategies for head-mounted magnification for low vision patients. *Transl Vis Sci Technol.* 2019;8(1):23.

15. Wilson E, Hilferty M. The distribution of chi-square. *Proc Natl Acad Sci USA.* 1931;17(12):684.

16. Massof R. A clinically meaningful theory of outcome measures in rehabilitation medicine. *J Appl Meas.* 2010;11(3):253–270.

17. Gobeille M, Bradley C, Goldstein J, Massof R. Activity Inventory calibrations: calibrated item measures and thresholds for the activity inventory. Available at: [https://sourceforge.net/projects/ai-calibrations/files/](https://sourceforge.net/projects/ai-calibrations/files/). Accessed April 16, 2021.