Published in final edited form as:

Optom Vis Sci. 2013 August ; 90(8): 788–798. doi:10.1097/OPX.0000000000000011.

Interpretation of Low Vision Rehabilitation Outcomes Measures

Robert W. Massof, PhD, FAAO and

Lions Vision Research and Rehabilitation Center, Wilmer Eye Institute, Johns Hopkins University School of Medicine. Baltimore, Maryland

Joan A. Stelmack, OD, MPH, FAAO

Blind Rehabilitation Center, Edward Hines Jr. Veterans Affairs Hospital, Hines, Illinois, Illinois Eye and Ear Infirmary, UIC Department of Ophthalmology and Visual Sciences, University of Illinois School of Medicine, Chicago, Illinois , and Illinois College of Optometry, Chicago, Illinois

Abstract

Purpose.—This paper presents a theoretical interpretation of patient-reported outcomes of low vision rehabilitation (LVR) employing rating scale questionnaires and uses previously published results of LVR outcome studies to illustrate theoretical points and validate assumptions.

Theory.—Patients' judgments of the difficulty they have performing tasks are interpreted as magnitude estimates of their functional reserve for each task, which is the difference between their visual ability and the visual ability demanded by the task. We assume that improvements in functional reserve can occur by increasing the patient's visual ability with medical, surgical, or refractive interventions or decreasing the visual ability demanded by the item with activity-specific vision assistive equipment, adaptations, and environmental modifications. Activity-specific interventions cause differential item functioning (intervention-related DIF). Intervention-related DIF makes the measured size of the treatment effect dependent on the item content and the mix of responsive and unresponsive items to intervention.

Conclusions.—Because Intervention-related DIF depends on the choice of items, the outcome measure selected should be appropriate to the aims of the intervention and the impairment level of the sample to demonstrate the full effects of an intervention. Items that are given extreme positive ratings at pre-intervention baseline (e.g., "not difficult") have no room for improvement. These items must also be filtered out because they will dilute the measured effect of the activity-specific interventions of LVR.

Keywords

low vision rehabilitation; patient-reported outcome measure; visual function questionnaire; differential item functioning; Rasch analysis; item response theory

Numerous low vision rehabilitation outcome studies have employed visual function rating scale questionnaires (VFQ) to measure the effectiveness of intervention.[1] Several of the earlier outcome studies employed traditional scoring methods that consisted of linear

Corresponding author: Robert W. Massof, Wilmer B-43, Johns Hopkins Hospital, 600 N. Wolfe St, Baltimore, Maryland 21287, bmassof@jhmi.edu.

combinations of ordinal response scores for all items in the VFQ, or subsets of items (i.e., domain scores).[2] More recent studies estimated interval scale outcome measures from VFQ item responses using Rasch models or, in some cases, item response theory (IRT) models.[3] Although the use of traditional scoring methods has been strongly criticized by Rasch and IRT model advocates,[4] if Rasch analysis confirms that valid measurements can be estimated from item responses for a particular instrument, then the estimated interval measure must be monotonic with the raw score (i.e., the sum of ordinal item response scores) and with any linear or other monotonic transformation of the raw score (i.e., a transformation that continuously, but not necessarily linearly, increases or decreases with increases or decreases in the raw score).[5] This monotonicity requirement of axiomatic measurement theory could be used to defend the choice of traditional scoring of VFQ responses in low vision rehabilitation outcome studies. However, traditional scoring still has serious weaknesses because the resulting scale is inherently nonlinear, instrument-specific, and distorted by missing item responses (which are permitted by most VFQs).[6]

A major strength of Rasch and IRT models is that the latent variables one is attempting to measure (i.e., variables that are not directly observed, but are inferred), and their relationship to the observed item responses, are explicitly defined.[7] One variable, called the *person measure*, quantifies the latent trait of interest, e.g., visual ability. Visual ability refers to the person’s ability to perform activities that depend on vision. Visual impairments are expected to modify a person’s visual ability. A second variable, the amount of visual ability required to perform a specific activity described by an item, is called the *item measure*. Item measures vary between items in a VFQ because the amount of visual ability required to perform different activities described by the items varies with the activity’s visual demand and how dependent successful performance of the activity is on vision (e.g., reading fine print in a contract has greater visual demand and depends more on vision than does buttoning a shirt).

Like the VA LV VFQ,[8–9] which was used as the primary outcome measure in the VA Low Vision Intervention Trial (LOVIT),[10] most VFQs ask respondents to rate the difficulty of performing each activity described by the items. The ordinal rating scale typically has four or five response options that range from “no difficulty” to “impossible to do”. Both Rasch and IRT models assume that when people rate the difficulty of performing an activity, they are judging the magnitude of the difference between their own visual ability and the amount of visual ability required to perform the activity – this difference is called *functional reserve*.[11] Respondents are likely to rate an activity as “not difficult “ if functional reserve is large, rate it as “very difficult” if functional reserve is close to, but still greater than, zero, and rate it as “impossible to do” if functional reserve is negative (i.e., the activity requires more visual ability than the person has).

If a person acquires a visual impairment, we expect her visual ability to be reduced and functional reserve to correspondingly be reduced by the same amount for every item. If the visual impairment is reversed as a result of clinical intervention, the person’s visual ability should improve and functional reserve should be increased by the same amount for every item. Thus, in the case of cataract surgery, anti-VEGF therapy, and other vision restoration procedures, person measures estimated from VFQ item responses, and the corresponding traditional scores, should improve.

Another way to increase functional reserve is to lower the amount of visual ability required to perform the activity described by each item. The visual ability required to perform an activity could be reduced by providing the patient with a vision assistive device, teaching the patient an adaptive strategy, modifying illumination, modifying the environment, or providing some other activity-specific intervention. Activity-specific interventions are expected to change item measures, not person measures. Changes in item measures with changes in properties of the respondents is called *differential item functioning* (DIF). In most circumstances, evidence of DIF is considered to be a very serious problem that indicates confounding variables are contaminating and distorting the measurement.[12] In those cases, every attempt is made to modify or remove items that exhibit DIF in order to minimize or eliminate the detrimental effects of DIF on the integrity of the measurement. However, in the case of low vision rehabilitation, we would view intervention-related DIF as a manifestation of a treatment effect. In the case of low vision rehabilitation, if we were to remove those items that exhibited intervention-related DIF, we would reduce the effects of treatment in our measurement, or even eliminate them completely.

More than a decade ago, we showed that the NEI VFQ-25 plus supplement exhibited strong evidence of intervention-related DIF when administered to patients before and after rehabilitation at a VA Blind Rehabilitation Center (BRC) and a VA VICTORS program.[13] We concluded that 7 of 34 items were responsive to intervention by exhibiting intervention-related DIF (although we did not use that term). There also was evidence of a small (0.5 logit) improvement in the person measure post-rehabilitation for BRC patients. That study motivated us to develop a new low vision outcome measure, the VA LV VFQ-48,[8] which employs items that describe activities targeted by blind and low vision rehabilitation programs in the VA centers. Outcome studies with the VA LV VFQ-48 showed that 7 of 48 items exhibited significant intervention-related DIF.[9] However, even though we did not succeed in eliminating intervention-related DIF, unlike the NEI VFQ-25, the VA LV VFQ-48 demonstrated large (1.5 logits) improvements in the person measure post-rehabilitation for BRC patients.

Demonstrations of an improvement in the person measure post-rehabilitation does not necessarily mean there was a change in the person trait. Changes in item responses as a consequence of intervention reflect changes in functional reserve, since that is what is being estimated with the difficulty ratings. If functional reserve improves for all items after activity-specific interventions, the smallest observed change, which is common to all items, could be assigned to the person measure and the balance of the changes in functional reserve could be assigned to the item measures in the explicit form of intervention-related DIF.

To be useful as an outcome measure, intervention-related DIF must be expressed as a single number that can be interpreted clinically, combined with changes in the person measure, and compared across patients and across services to evaluate the effectiveness of rehabilitation on a univariate scale. The aim of this paper is to present a theoretical interpretation of intervention-related DIF and demonstrate how it can be used to better understand the relative contributions of item measure changes and person measure changes to a univariate measurement of outcomes for low vision rehabilitation.

THEORY

Rasch Model

The person measure is represented by a variable named P and the item measure is represented by a variable named I. We use the subscript n as an index for the person and the subscript j as an index for the item, i.e. P_n is the person measure for person n and I_j is the item measure for item j. The functional reserve for person n encountering item j is $F_{nj} = P_n - I_j$. The VA LV VFQ has four difficulty response categories: 1) not difficult, 2) slightly or moderately difficult, 3) very difficult, and 4) impossible to do. Person n is asked to report the difficulty he experiences with the activity described by item j. We assume the person is judging the magnitude of the functional reserve he has for item j and uses his own response criteria for choosing a difficulty category. If C_{n1}, C_{n2}, and C_{n3} are the response criteria of person n, then theoretically we represent the person's response rule as:

- Respond "impossible to do" if $F_{nj} < C_{n1}$
- Respond "very difficult" if $C_{n1} < F_{nj} < C_{n2}$
- Respond "slightly or moderately difficult" if $C_{n2} < F_{nj} < C_{n3}$
- Respond "not difficult" if $F_{nj} > C_{n3}$.

The average response criteria for the population of low vision patients are C_1, C_2, and C_3. We assume that each person differs from the average criterion C_x by the amount e_{nx}, i.e., $C_{nx} = C_x + e_{nx}$ for x equal to 1, 2, or 3 (e_{nx} can be thought of as the error introduced by using the average criterion for response category x rather than using person n's personal criterion). We further assume that the item measure, I_j is the average for the population. Different people might interpret the item differently and the conditions under which the activity described by the item is performed might vary between people. Thus, for a given person, functional reserve should be defined as $F_{nj} = P_n - I_j - e_{nj}$ (similarly, e_{nj} can be thought of as the error for person n introduced by employing the average item measure for item j). Because e_{nj} can be added to all of the terms in the decision rule without changing the rule, we can replace F_{nj} with $P_n - I_j$ and C_{nx} with $C_x + e_{njx}$, where $e_{njx} = e_{nx} + e_{nj}$ (we can think of e_{njx} as the total error introduced by employing the average item measure for item j and the average criterion for response category x rather than the personal values). By definition, the average value of e_{njx} across all people in the population must be zero and, it probably is safe to assume, e_{njx} is normally distributed across people.

If we add the requirements that the standard deviation of the distribution of e_{njx} values across people is the same for every item and response criterion and that there are no correlations in e_{njx} values between people (across items and response criteria), between items (across people and response criteria), or between response criteria (across people and items), then we have described a Rasch model (IRT models make different assumptions about the statistics of e_{njx}; see reference 14 for a more detailed mathematical description of Rasch and IRT models). By making these assumptions about the statistics and distribution of

e_{njx}, we are able to employ maximum likelihood estimation methods to estimate the person measure for every person, the average item measure for every item and the average response criteria for all response categories from a data set of item responses made by a sample of observers representing the population of interest. *Rasch analysis* is employed to test the validity of the Rasch model's statistical assumptions for the sample of persons, items, and response categories tested (see reference 14 for an overview of Rasch analysis and reference 15 for lessons on Rasch analysis).

Application of the Rasch Model to Interpreting Outcome Measures

It is important to note that the same unit of measurement is used for person measures, item measures (and therefore functional reserve), and response category criteria. Because the person is judging the magnitude of functional reserve, an increase in the person measure will have the same effect as a decrease in the item measure, or vice versa.

In the case of an outcome study, we can represent the person and item measures as functions of time: $P_n(t)$ for person n and $I_j(t)$ for item j. For each person in the outcome study, the clock starts with the pre-intervention baseline measurement. Therefore, we will designate the time of the baseline measurement as t_0. The first post-intervention follow-up measurement occurs at some later time that we will designate as t_1. Thus, estimated functional reserve at baseline is $F_{nj}(t_0) = P_n(t_0) - I_j(t_0)$ and functional reserve at the first follow-up measure is $F_{nj}(t_1) = P_n(t_1) - I_{nj}(t_1)$, for person n and item j. In the latter case, the item measure for person n at follow-up, i.e., $I_{nj}(t_1)$, does not just refer to the inclusion of the random departure from the average item measure (the random effect), as discussed above, it also includes the activity-specific effect of intervention for that person (the fixed effect). The change in person measure from baseline to the first follow-up is $\Delta P_n = P_n(t_1) - P_n(t_0)$ and the change in item measure for person n over the same interval is $\Delta I_{nj} = I_{nj}(t_1) - I_j(t_0)$. Thus, we can rewrite the equation for functional reserve at follow-up as $F_{nj}(t_1) = P_n(t_0) + \Delta P_n - I_j(t_0) - \Delta I_{nj}$ and the effect of intervention on functional reserve is $\Delta F_{nj} = F_{nj}(t_1) - F_{nj}(t_0) = \Delta P_n - \Delta I_{nj}$, i.e., the change in functional reserve for person n and item j is the difference between the change in the person measure and the change in the item measure.

The next step is to summarize the effects of intervention with a single outcome variable. But, before proceeding with that task, let us return to the baseline measures and demonstrate a useful approximation to the maximum likelihood method of estimating person measures. First, we start with a set of item responses at pre-intervention baseline for our representative sample of persons and use the maximum likelihood procedure to estimate person measures for every person, average item measures for every item, and average response criteria for the response categories. Second, using the estimated person and item measures, we calculate functional reserve for every item paired with every person, i.e., $F_{nj}(t_0) = P_n(t_0) - I_j(t_0)$. Third, we average functional reserve for all person/item pairs for which the observed response was "not difficult". We similarly calculate average functional reserve for each of the other response categories. This averaging operation is

$\bar{F}_x(t_0) = \Sigma_n \Sigma_j \frac{P_n(t_0) - I_j(t_0)}{(NJ)_x} = \bar{P}_x(t_0) - \bar{I}_x(t_0)$ for *(NJ)*$_x$ person/item pairs for which the difficulty rating was *x*. This step enables us to translate each difficulty rating to average functional reserve (a.k.a. average measure[16]). Fourth, we approximate average functional reserve for difficulty rating *x* with an idealized definition of average functional reserve for difficulty rating *x* by person *n* to item j, $\bar{F}_{x(j)} \cong \hat{P}_{nj}(x) - I_j(t_0)$ where $I_j(t_0)$ is the baseline measure for item *j* and $\hat{P}_{nj}(x)$ is the estimate of the measure for person *n* from his response of *x* to item *j*. From this approximation, we obtain a separate estimate of the person measure for each item response, $\hat{P}_{nj}(x) \cong \bar{F}_{x(j)} + I_j(t_0)$. Finally, we average these person measure estimates across all items, $\hat{P}_n = \sum_{j=1}^{J} \frac{\hat{P}_{nj}(x)}{J}$.

Figure 1 graphically illustrates in a keyform plot the reading function person measure estimate from average functional reserve values for one of the LOVIT participants. The reading function person measure scale is plotted on the horizontal axis and each row corresponds to a different reading item in the VA LV VFQ-48. The item measures at baseline are listed next to the item descriptions on the right side of the figure. Each symbol corresponds to a different difficulty rating. The table at the bottom of the figure lists the average functional reserve computed for each response category across all persons and items at pre-intervention baseline. The symbols for each item are positioned horizontally on the person measure scale according to the sum of average functional reserve and the baseline item measure. The lines connect the participant's responses to the items. The vertical line is the average horizontal position of the symbols that correspond to the participant's item responses, which is the final estimate of the person measure for the participant.

Figure 2 illustrates a scatter plot of this keyform-type estimate of the person measure versus the person measure estimated using the standard maximum likelihood method for a combined sample of outcome study participants[9] and LOVIT participants[10] at pre-intervention baseline. The approximation fails in the tails because the average functional reserve for the extreme response categories (i.e., not difficult and impossible to do) impose a ceiling and a floor on the person measure approximations. The approximation can be improved by correcting for the ceiling and floor effects with an inverse hyperbolic tangent transform of the estimated person measures.[17] The results illustrate the validity of person measure estimates approximated from average functional reserve.

Analysis of Intervention-Related Differential Item Functioning (DIF)

For interventions that aim to restore the patient's vision, we expect that intervention-related changes in functional reserve for each item will be the consequence of changes in the person measure, i.e., $\Delta F_{nj} = \Delta P_n$, with no change in the item measure, i.e., $\Delta I_{nj} = 0$. More generally, in terms of the approximation from average functional reserve,

$$\Delta\hat{P}_n = \hat{P}_n(t_1) - \hat{P}_n(t_0) \cong \sum_{j=1}^{J} \frac{F_{nx(j)}(t_1) + I_{nj}(t_1) - F_{nx(j)}(t_0) - I_{nj}(t_0)}{J} = \sum_{j=1}^{J} \frac{\Delta F_{nx(j)} + \Delta I_{nj}}{J}$$

, which, if $\Delta I_{nj} = 0$ for all items, reduces to $\Delta\hat{P}_n \cong \overline{\Delta F}_n$. In the case of intervention-related

DIF, $\Delta I_{nj} \neq 0$ for all items and $\Delta \hat{P}_n \cong \sum_{j=1}^{J} \frac{\Delta F_{nx(j)} + \Delta I_{nj}}{J} = \overline{\Delta F}_n + \sum_{j=1}^{J} \frac{\Delta I_{nj}}{J}$. At the opposite extreme, if all intervention is activity-specific and has no effect on the person measure, i.e., $\Delta P_n = 0$, then $\overline{\Delta F}_n = -\sum_{j=1}^{J} \frac{\Delta I_{nj}}{J}$. However, if all item measures change between the two time points by at least a minimum of ΔI_m, then we can define the change in item measure j for person n as $\Delta I_{nj} = \Delta I'_{nj} + \Delta I_{nm}$, where $\Delta I'_{nj}$ is the residual DIF. Thus, even if the person measure does not change as a result of intervention, the average change in functional reserve would be $\overline{\Delta F}_n = -\sum_{j=1}^{J} \frac{\Delta I'_{nj} + \Delta I_{nm}}{J} = -\Delta I_{nm} - \sum_{j=1}^{J} \frac{\Delta I'_{nj}}{J}$, which is indistinguishable from $\overline{\Delta F}_n = \Delta P_n - \sum_{j=1}^{J} \frac{\Delta I'_{nj}}{J}$.

One analytic strategy is to estimate person measures with the follow-up item measures anchored to the baseline item measures. With that approach, we impose the constraint that $\Delta I_{nj} = 0$ and by definition the estimated change in person measure is equal to the average change in functional reserve across items. However, we understand that the average change in functional reserve actually is the difference between the real change in person measure and minimum change in all item measures minus the average residual DIF across all items, i.e., $\overline{\Delta F}_n = \Delta P_n - \Delta I_{nm} - \sum_{j=1}^{J} \frac{\Delta I'nj}{J}$.

To demonstrate the effect of intervention-related DIF, Figure 3 illustrates the same keyform plot illustrated in Figure 1 except that we have added the VA LV VFQ-48 reading item difficulty ratings of the same LOVIT participant at post-low vision rehabilitation follow up (connected by the dashed lines). The difficulty ratings for two of the items, “Read newspaper headlines” and “Read menus”, were the same at the post-rehabilitation follow-up as they were at the pre-rehabilitation baseline (highlighted with large circles). The dashed vertical line is the average position on the reading function person measure scale of the sum of average functional reserve, corresponding to the difficulty rating, and the baseline item measure for each item. Because two of the item responses did not change after rehabilitation and all the others did, we interpret these results as evidence of intervention-related DIF. It appears that intervention was ineffective at increasing this participant’s functional reserve for reading newspaper headlines and reading menus, but was effective to differing degrees in increasing functional reserve for the other reading tasks.

Intervention-Related DIF and Effect Size

Figure 4 compares the results of several low vision rehabilitation outcome studies, and groups within studies, that were reported in the literature over the past 15 years.[2,3,9,10,18–28] For each study, outcomes were assessed with a VFQ administered at pre-intervention baseline and again at a post-intervention follow-up. Different studies employed different VFQs and different follow-up times. Some studies reported VFQ results using traditional scoring based on sums of ordinal item scores and other studies employed interval scale measures estimated using Rasch models. To equate scales, outcome measures for all studies were transformed to effect sizes (i.e., Cohen’s d),[29] which is the difference between the mean VFQ score across all patients at post-intervention follow-up and the mean VFQ score across all patients at pre-intervention baseline divided by the standard deviation of the VFQ scores at baseline. Effect sizes for most studies are small (<0.3) or moderate (≅0.5), but they are very large for the VA Blind Rehabilitation Center (2.1) and for LOVIT (2.5). This large discrepancy in effect sizes says that the VA programs are exceptionally effective and private sector programs are not, and/or the VFQs employed by the different studies are measuring different patient traits or have different levels of responsiveness to the effects of intervention.

It is unlikely that the different VFQs measured different patient traits. The low vision participants in the different studies were very similar and the items in the various questionnaires described similar everyday activities. Past studies in which different VFQs were administered to the same low vision patients demonstrated that the estimated person measures from very different VFQs are all in agreement.[30–31]

The VA LV VFQ-48 was administered to VA Blind Rehabilitation Center patients[9] and to LOVIT participants,[10] both of whom showed very large effects of intervention (d=2.1 and 2.5 respectively), and it was administered to patients in two private outpatient low vision clinics,[9] both of whom showed small to moderate effects of intervention (d = 0.36 and 0.43). Also, Pearce et al.[26] observed a larger 1-month effect for a low vision service that included an appointment with a hospital optometrist and a visit with an optician who provided instruction and assistance with prescribed low vision devices (d=0.70) than for a low vision service that only included a single appointment with a hospital optometrist (d=0.55) when measuring outcomes with the Activity Inventory[31] and Court et al.[27] observed a larger effect (d=0.46) for hospital-based (comprehensive) then for community-based (primary, d=0.29) low vision services when using a subset of NEI VFQ-25 plus supplement items for the outcome measure. These results suggest the possibility of very real differences in the effectiveness of different clinical programs. But, there also appears to be a difference in responsiveness of different instruments – the effect size for the VA Blind Rehabilitation Center was 30% smaller when the NEI VFQ-25 plus supplement was employed (d = 1.5).[13]

One conclusion drawn from a secondary analysis of LOVIT results was that the only predictor of the magnitude of change in person measure post-rehabilitation was the magnitude of the baseline person measure.[10] Large change scores were associated with low visual ability at baseline and smaller change scores were associated with higher visual ability at baseline. Figure 5 is a scatter plot of the change in person measure post-intervention versus baseline person measures, both estimated from VA LV VFQ-48 responses of VA Blind Rehabilitation Center patients (filled circles)[9] and private outpatient low vision clinic patients (open circles).[9] The linear trend in the data confirms the LOVIT conclusion that the magnitude of the measured effect of rehabilitation is associated with the magnitude of the person measure at baseline.

All of these varied and seemingly contradictory outcome measure results can be explained by assuming that the major effects of low vision rehabilitation are item-specific and variations in effect sizes for different VFQs and different baseline levels of visual ability are manifestations of intervention-related DIF. In the case of traditional scoring, the ordinal item scores are monotonic with functional reserve; thus, the person raw score, or any linear transformation of the raw score, is monotonic with average functional reserve across items. In the case of Rasch models, we employ the approximation of the change in person measure estimate based on the average change in functional reserve, the largest change in item measure shared by all items, and the average residual DIF introduced earlier, i.e., $\Delta\hat{P}_n = \overline{\Delta F}_n + \Delta I_{nm} + \sum_{j=1}^{J} \frac{\Delta I'_{nj}}{J}$. If the item measure does not change for person n on one or more items in the VFQ, then $\Delta I_{nm} = 0$ and if $\Delta P_n = 0$ then $\Delta\hat{P}_n$ is determined entirely by average intervention-related DIF.

The dependence of effect size on baseline visual ability that is demonstrated in Figure 5 could be a consequence of intervention-related DIF and a bounded rating-scale. Patients with greater visual ability at baseline are likely to rate more items at baseline as "not difficult" than do patients with lesser visual ability at baseline. If L out of J item responses are "not difficult" at baseline, there is no room for improvement at follow-up on these items; only $K = J - L$ items can exhibit improvement. Thus, if intervention causes a change in item difficulty of ΔI_{nk} for person n in the K items that have room to change, and no change in the $J - K$ items that are "not difficult" at baseline, then the estimated change in person measure is $\Delta\hat{P}_n = \overline{\Delta F}_n + \sum_{j=1}^{J} \frac{\Delta I_{nj}}{J} = \overline{\Delta F}_n + \sum_{k=1}^{K} \frac{\Delta I_{nk}}{K+L}$, where the sum is divided by $J = K + L$ because we are averaging over all items and $\Delta I_{nj} = 0$ for $j \neq k$.

An example of responses to nine reading items by a low vision patient served by a private outpatient center[9] is shown in the keyform plot in Figure 6. The person measure estimates from responses at baseline are connected by solid lines and the person measure estimates from responses at post-intervention follow-up are connected by dashed lines. The solid vertical line is the average person measure estimate at baseline over the nine items and the dashed vertical line is the average person measure estimate over the same nine items at post-intervention follow-up. The difference between the two vertical lines is the average change in the estimated person measure, i.e., $\Delta\hat{P} = 0.76$ logit. Note that the baseline response is "not difficult" (open circles) for five items: "Read newspaper headlines", "Read newspaper or magazine articles", "Read mail", "Read menus", and "Read signs". These items also were rated "not difficult" at post-intervention follow-up.

Figure 7 illustrates the same keyform plot shown in Figure 6, except that the five items that were rated "not difficult" at baseline have been removed. The solid vertical line is the average estimated person measure at baseline from the responses to the four retained items and the dashed vertical line is the average estimated person measure at post-intervention follow-up from responses to the same four retained items. The difference between the lines is the average change in the estimated person measure for the five items that were rated as having some non-zero level of difficulty at baseline, i.e., $\Delta\hat{P} = 1.72$. For the data in Figure 6, the estimated change in person measure is $\Delta\hat{P}_n = \overline{\Delta F}_n + \sum_{k=1}^{K} \frac{\Delta I_{nj}}{K+L}$ with $K + L = 9$. For the data in Figure 7, the estimated change in person measure is $\Delta\hat{P}_n = \overline{\Delta F}_n + \sum_{k=1}^{K} \frac{\Delta I_{nk}}{K}$ with $K = 4$. These examples demonstrate that the measured intervention effect size varies with the responsiveness of items when the intervention is activity-specific. That is, the inclusion of items in the outcome measure that do not require intervention or are unresponsive to intervention can reduce the measured effect size.

DISCUSSION

If a low vision patient identifies an activity at the baseline evaluation as “not difficult”, “not applicable”, or “not important” to them, that activity most likely would not be included as a low vision rehabilitation goal in the plan of treatment. Most VFQs, including the VA LV VFQ, offer patients the opportunity to opt out of rating an item by responding that they “do not do for reasons other than vision” or that the item is “not applicable” to them. This type of opt-out response functions as an item filter[32] because it is scored as missing data. Technically, traditional scoring based on linear transformations of raw scores cannot tolerate missing data and must impute the missing values.[33] An advantage of estimating interval-scaled person measures from item responses using Rasch or IRT models is that missing data mainly affect the precision of the estimate and have less of an effect on the accuracy (albeit accuracy can be affected if missing data change the ceiling or the floor of the estimated measures, as occurs in the example in Figure 7).

A more important issue for outcome measures is that items included in a VFQ that describe activities that will not be addressed by low vision rehabilitation, and therefore would not be expected to exhibit a change in functional reserve after rehabilitation, will reduce measured effectiveness because intervention-related DIF is averaged across items. Items not exhibiting DIF do not contribute to the numerator, but do contribute to the denominator in calculating the average, thereby reducing the estimated change in functional ability. Thus, item content in low vision rehabilitation outcome measures must be chosen carefully to produce clinically meaningful measures. To this end, Ryan and her colleagues suggested using only the seven most responsive items in the NEI VFQ-25 to measure low vision rehabilitation outcomes.[34] The VA LV VFQ incorporates items that describe activities that are important and difficult for most legally blind VA low vision patients and are addressed by the VA blind and low vision rehabilitation programs for all eligible patients. However, in addition to serving legally blind veterans, new outpatient low vision rehabilitation clinics recently opened in the VA health care system can serve low vision patients who are not legally blind and for whom many of the items in the VA LV VFQ are not difficult at baseline, which could lead to underestimates of the effectiveness of intervention.

It is likely that the relatively small effect sizes seen in low vision rehabilitation outcome studies performed on non-VA patients reflect a combination of real differences in the effectiveness of different programs and, because of intervention-related DIF, differences between patient samples in the baseline level of functional ability and differences between VFQs in item content and their responsiveness to activity-specific effects of intervention. Thus, it appears that intervention-related DIF makes it possible to heavily influence measured outcomes with item selection and patient eligibility criteria. So, how do we prevent chaos in clinical research on low vision rehabilitation?

A clinically meaningful low vision outcome measure should be appropriate to the aims of the intervention and the impairment level of the sample. Some components of low vision rehabilitation might change the patient's vision, e.g., refractive error correction, but most are targeted at making vision-dependent activities easier to perform by using vision assistive equipment (e.g., magnifying devices, special lighting), adaptations (e.g., non-visual strategies to accomplish vision-dependent activity goals), environmental modifications (e.g., to increase visibility and/or safety), visual skills training (e.g., viewing, scanning, and perceptual motor skills training to improve acquisition and interpretation of visual information), and counseling (e.g., to help the patient revalue goals and cope with losses of functional capabilities). The effects of interventions that aim to improve function (e.g., orientation and mobility instruction), are likely to have effects on a large number of related activities, but the magnitude of the effect might be different for different activities. Interventions that aim to improve the performance of instrumental activities of daily living (e.g., adaptations and vision assistive equipment) are likely to produce effects that are activity-specific and must be assessed on an activity-by-activity basis.

In many cases, low vision clinicians develop an individualized plan of treatment based on the patient's functional history, the clinician's estimate of the patient's rehabilitation potential, and resources available to the patient.[35–36] In these cases, intervention-related DIF might require the use of adaptive VFQs[31,37] that customize the items included in the measure to match the preferences of individual patients and/or comport with their plans of treatment. At the very least, intervention-related DIF will require filtering of items rated not difficult at baseline in order to avoid diluting the effects of intervention by including items that have no room for improvement.

ACKNOWLEDGMENTS

This study was supported by grants from the National Eye Institute (EY022322) and the Department of Veterans Affairs Rehabilitation Research and Development Service (C3457).

REFERENCES

1. Binns AM, Bunce C, Dickinson C, Harper R, Tudor-Edwards R, Woodhouse M, Linck P, Suttie A, Jackson J, Lindsay J, Wolffsohn J, Hughes L, Margrain TH. How effective is low vision service provision? A systematic review. Surv Ophthalmol 2012;57:34–65. [PubMed: 22018676]
2. Scott IU, Smiddy WE, Schiffman J, Feuer WJ, Pappas CJ. Quality of life of low-vision patients and the impact of low-vision services. Am J Ophthalmol 1999;128:54–62. [PubMed: 10482094]
3. Smith HJ, Dickinson CM, Cacho I, Reeves BC, Harper RA. A randomized controlled trial to determine the effectiveness of prism spectacles for patients with age-related macular degeneration. Arch Ophthalmol 2005;123:1042–50. [PubMed: 16087836]
4. Lamoureux E, Pesudovs K. Vision-specific quality-of-life research: a need to improve the quality. Am J Ophthalmol 2011;151:195–7. [PubMed: 21251493]
5. Las Hayas C, Bilbao A, Quintana JM, Garcia S, Lafuente I. A comparison of standard scoring versus Rasch scoring of the visual function index-14 in patients with cataracts. Invest Ophthalmol Vis Sci 2011;52:4800–7. [PubMed: 21527383]
6. Massof RW. Likert and Guttman scaling of visual function rating scale questionnaires. Ophthalmic Epidemiol 2004;11:381–99. [PubMed: 15590585]
7. Massof RW. The measurement of vision disability. Optom Vis Sci 2002;79:516–52. [PubMed: 12199545]

8. Stelmack JA, Szlyk JP, Stelmack TR, Demers-Turco P, Williams RT, Moran D, Massof RW. Psychometric properties of the Veterans Affairs Low-Vision Visual Functioning Questionnaire. Invest Ophthalmol Vis Sci 2004;45:3919–28. [PubMed: 15505037]

9. Stelmack JA, Szlyk JP, Stelmack TR, Demers-Turco P, Williams RT, Moran D, Massof RW. Measuring outcomes of vision rehabilitation with the Veterans Affairs Low Vision Visual Functioning Questionnaire. Invest Ophthalmol Vis Sci 2006;47:3253–61. [PubMed: 16877389]

10. Stelmack JA, Tang XC, Reda DJ, Rinne S, Mancil RM, Massof RW. Outcomes of the Veterans Affairs Low Vision Intervention Trial (LOVIT). Arch Ophthalmol 2008;126:608–17. [PubMed: 18474769]

11. Massof RW. A systems model for low vision rehabilitation. II. Measurement of vision disabilities. Optom Vis Sci 1998;75:349–73. [PubMed: 9624700]

12. Zumbo BD. Three generations of DIF analyses: considering where it has been, where it is now, and where it is going. Lang Assess Q 2007;4:223–33.

13. Stelmack JA, Stelmack TR, Massof RW. Measuring low-vision rehabilitation outcomes with the NEI VFQ-25. Invest Ophthalmol Vis Sci 2002;43:2859–68. [PubMed: 12202503]

14. Massof RW. 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;18:1–19. [PubMed: 21275592]

15. Bond TG, Fox CM. Applying the Rasch Model: Fundamental Measurement in the Human Sciences, 2nd ed. Mahwah, NJ: L. Erlbaum Associates; 2007.nd

16. Linacre JM. Investigating rating scale category utility. J Outcome Meas 1999;3:103–22. [PubMed: 10204322]

17. Massof RW. Application of stochastic measurement models to visual function rating scale questionnaires. Ophthalmic Epidemiol 2005;12:103–24. [PubMed: 16019693]

18. Wolffsohn JS, Cochrane AL. Design of the low vision quality-of-life questionnaire (LVQOL) and measuring the outcome of low-vision rehabilitation. Am J Ophthalmol 2000;130:793–802. [PubMed: 11124300]

19. McCabe P, Nason F, Demers Turco P, Friedman D, Seddon JM. Evaluating the effectiveness of a vision rehabilitation intervention using an objective and subjective measure of functional performance. Ophthalmic Epidemiol 2000;7:259–70. [PubMed: 11262673]

20. Hinds A, Sinclair A, Park J, Suttie A, Paterson H, Macdonald M. Impact of an interdisciplinary low vision service on the quality of life of low vision patients. Br J Ophthalmol 2003;87:1391–6. [PubMed: 14609841]

21. Reeves BC, Harper RA, Russell WB. Enhanced low vision rehabilitation for people with age related macular degeneration: a randomised controlled trial. Br J Ophthalmol 2004;88:1443–9. [PubMed: 15489491]

22. La Grow SJ. The effectiveness of comprehensive low vision services for older persons with visual impairments in New Zealand. J Visual Impair Blin 2004;98:679–92.

23. de Boer MR, Twisk J, Moll AC, Volker-Dieben HJ, de Vet HC, van Rens GH. Outcomes of low-vision services using optometric and multidisciplinary approaches: a non-randomized comparison. Ophthalmic Physiol Opt 2006;26:535–44. [PubMed: 17040417]

24. Lamoureux EL, Pallant JF, Pesudovs K, Rees G, Hassell JB, Keeffe JE. The effectiveness of low-vision rehabilitation on participation in daily living and quality of life. Invest Ophthalmol Vis Sci 2007;48:1476–82. [PubMed: 17389474]

25. Walter C, Althouse R, Humble H, Smith W, Odom JV. Vision rehabilitation: recipients' perceived efficacy of rehabilitation. Ophthalmic Epidemiol 2007;14:103–11. [PubMed: 17613844]

26. Pearce E, Crossland MD, Rubin GS. The efficacy of low vision device training in a hospital-based low vision clinic. Br J Ophthalmol 2011;95:105–8. [PubMed: 20837788]

27. Court H, Ryan B, Bunce C, Margrain TH. How effective is the new community-based Welsh low vision service? Br J Ophthalmol 2011;95:178–84. [PubMed: 20601662]

28. Wang BZ, Pesudovs K, Keane MC, Daly A, Chen CS. Evaluating the effectiveness of multidisciplinary low-vision rehabilitation. Optom Vis Sci 2012;89:1399–408. [PubMed: 22902419]

29. Cohen J Statistical Power Analysis for the Behavioral Sciences, 2nd ed. Hillsdale, NJ: L. Erlbaum Associates; 1988.

30. Massof RW. An interval-scaled scoring algorithm for visual function questionnaires. Optom Vis Sci 2007;84:689–704.

31. Massof RW, Ahmadian L, Grover LL, Deremeik JT, Goldstein JE, Rainey C, Epstein C, Barnett GD. The Activity Inventory: an adaptive visual function questionnaire. Optom Vis Sci 2007;84:763–74. [PubMed: 17700339]

32. Reardon SF, Raudenbush SW. A partial independence item response model for surveys with filter questions. Sociol Methodol 2006;36:257–300.

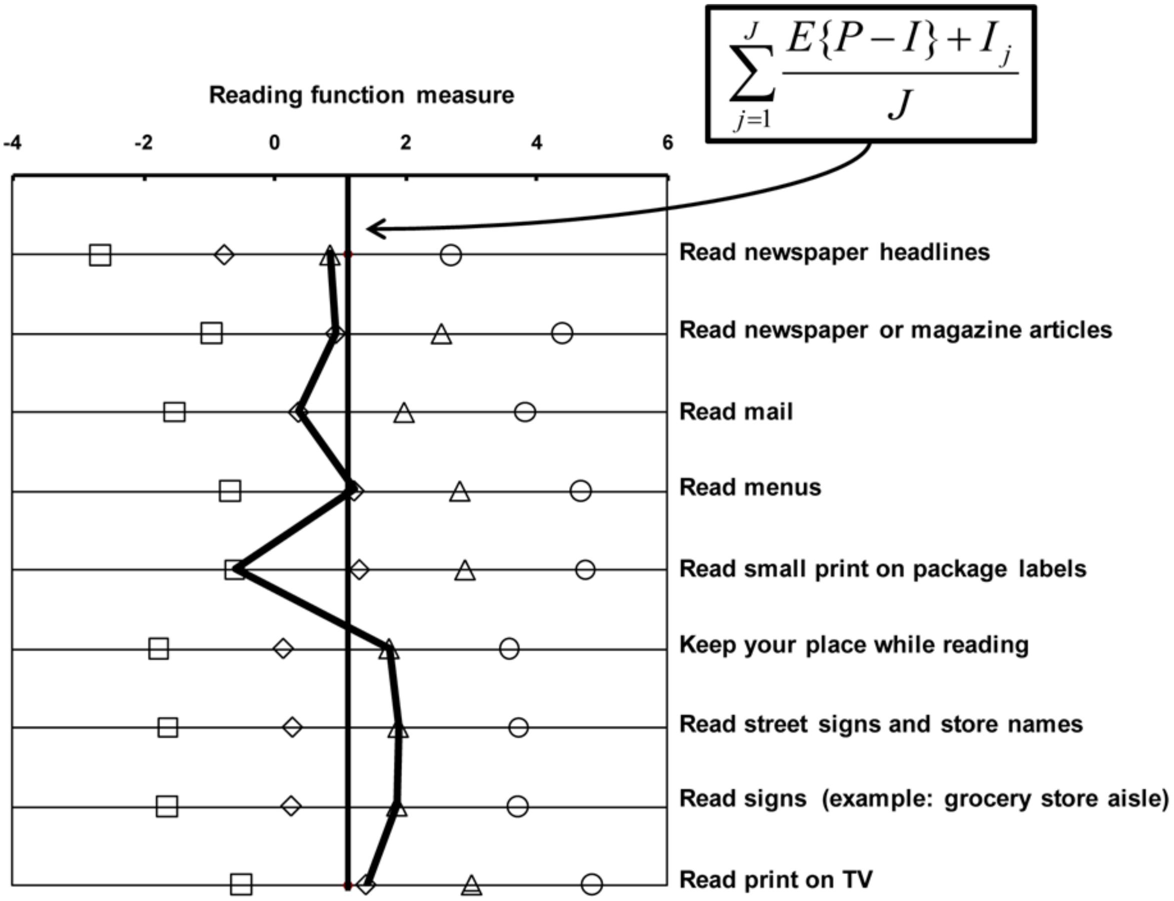
33. Fayers PM, Curran D, Machin D. Incomplete quality of life data in randomized trials: missing items. Stat Med 1998;17:679–96. [PubMed: 9549816]

34. Ryan B, Court H, Margrain TH. Measuring low vision service outcomes: Rasch analysis of the seven-item National Eye Institute Visual Function Questionnaire. Optom Vis Sci 2008;85:112–21. [PubMed: 18296928]

35. Massof RW. A systems model for low vision rehabilitation. I. Basic concepts. Optom Vis Sci 1995;72:725–36. [PubMed: 8570162]

36. Markowitz SN. Principles of modern low vision rehabilitation. Can J Ophthalmol 2006;41:289–312. [PubMed: 16767184]

37. Bruijning J, van Nispen R, Knol D, van Rens G. Low vision rehabilitation plans comparing two intake methods. Optom Vis Sci 2012;89:203–14. [PubMed: 22198794]



$$\sum_{j=1}^{J}\frac{E\{P-I\}+I_j}{J}$$

I_j
0.02
1.72
1.15
2
2.08
0.92
1.06
1.04
2.18

Response	Symbol	E{P – I}
Not difficult	○	2.65
Slightly/Moderately difficult	△	0.82
Very difficult	◇	-0.79
Impossible to do	□	-2.68

Figure 1.

Keyform plot of pre-intervention baseline responses to reading items in the VA LV VFQ by a participant in LOVIT. Reading function person measure is plotted on the horizontal axis at the top of the figure. Each of the four difficulty ratings is represented by a different symbol identified in the table at the bottom of the figure. That table also lists average functional reserve for each response category. Each row in the figure corresponds to a different reading item identified by the item labels on the right of the figure. The item measures are listed in the table to the right of the item descriptions. The response symbols are positioned on the horizontal axis according to the sum of functional reserve and the item measure. The participant’s responses to the reading items are connected by the solid black lines. The solid vertical line corresponds to the average of the sums of average functional reserve and the item measures, which mathematically is an estimate of the person measure.

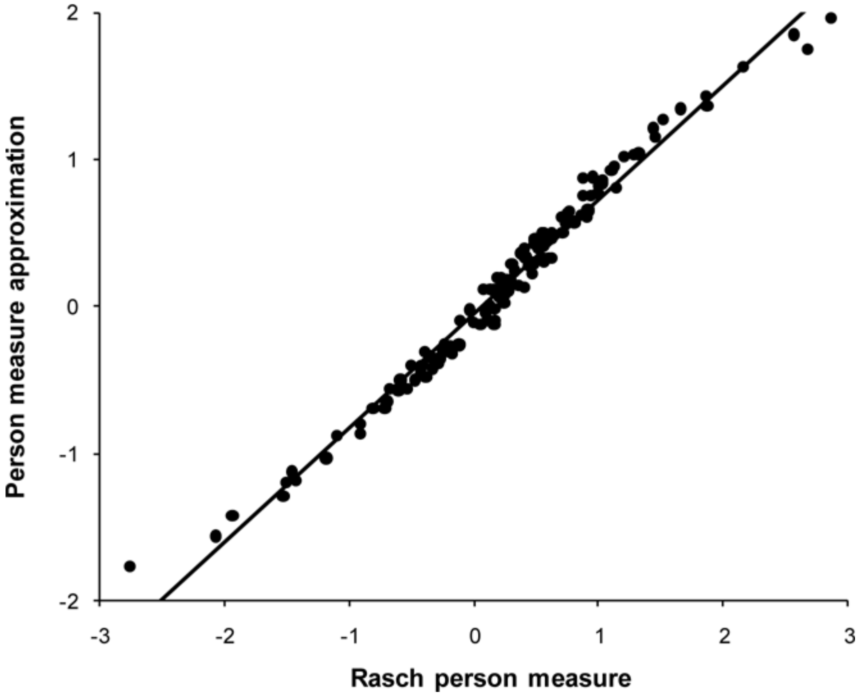


Figure 2.

Scatter plot of person measures estimated from average functional reserve (vertical axis) compared to person measures estimated using maximum likelihood estimation procedures with the Rasch model (horizontal axis) for LOVIT participants at pre-intervention baseline. There is strong agreement between the two sets of estimated measures (r^2 = 0.98).

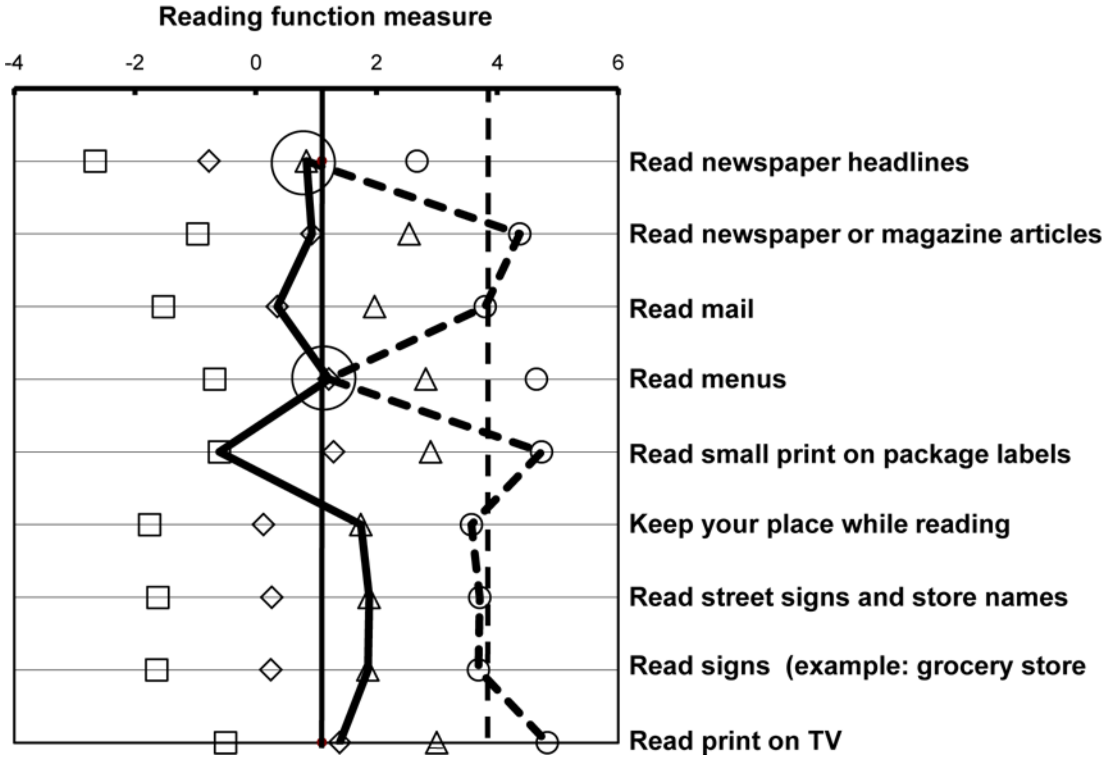


Figure 3.

Keyform plot for the same LOVIT participant whose responses are shown in Figure 1. The solid lines connect the participant's difficulty ratings at baseline and the solid vertical line is the average person measure estimate across all reading items (reproduction of the results shown in Figure 1). The dashed lines connect the same participant's difficulty ratings at post-intervention follow-up. For two items, "Read newspaper headlines" and "Read menus", the difficulty rating at follow-up was the same as the difficulty rating at baseline (highlighted with large circles). The difficulty ratings for all other items at follow-up were lower than they were at baseline. The dashed vertical line is the average of the person measure estimates for each item across all nine reading items.

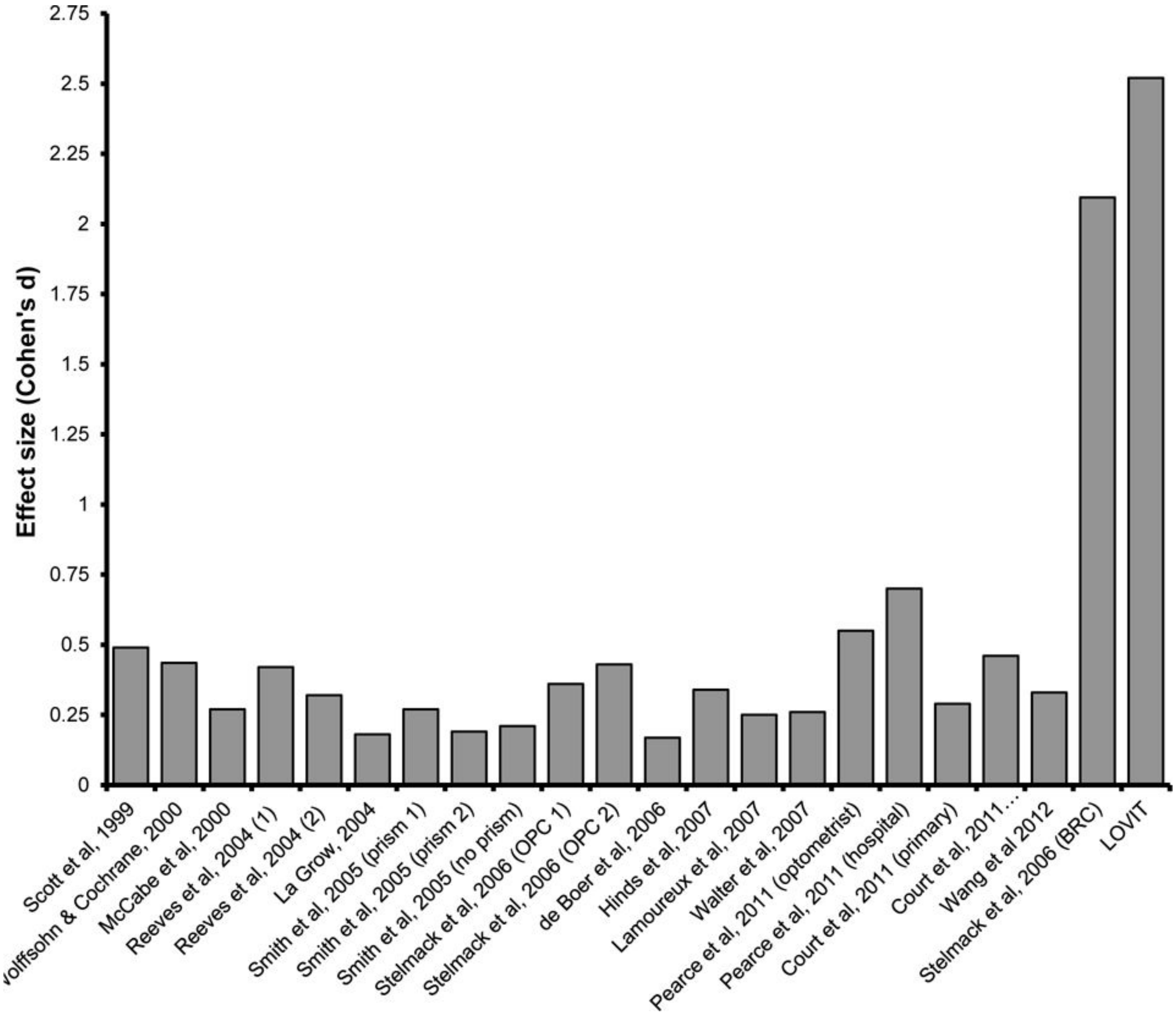


Figure 4.

Comparison of effect sizes for different low vision rehabilitation outcome studies and groups of patients within studies. Effect size is expressed as the difference between the mean outcome score at post-intervention follow-up and the score at pre-intervention baseline divided by the standard deviation of the baseline score distribution. Different studies employed different VFQs as the outcome measure and different methods of scoring.

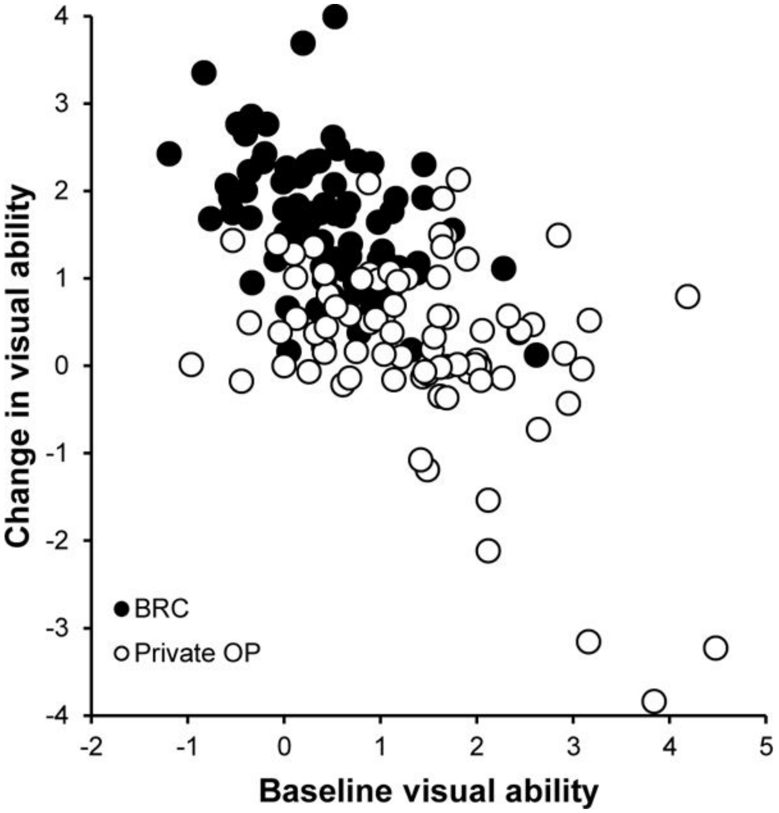


Figure 5.

Scatter plot of changes in visual ability (i.e., visual ability measured at post-intervention follow-up minus visual ability measured at baseline) versus visual ability measured at baseline for VA Blind Rehabilitation Center inpatients (filled circles) and for outpatients at private low vision rehabilitation centers (open circles).

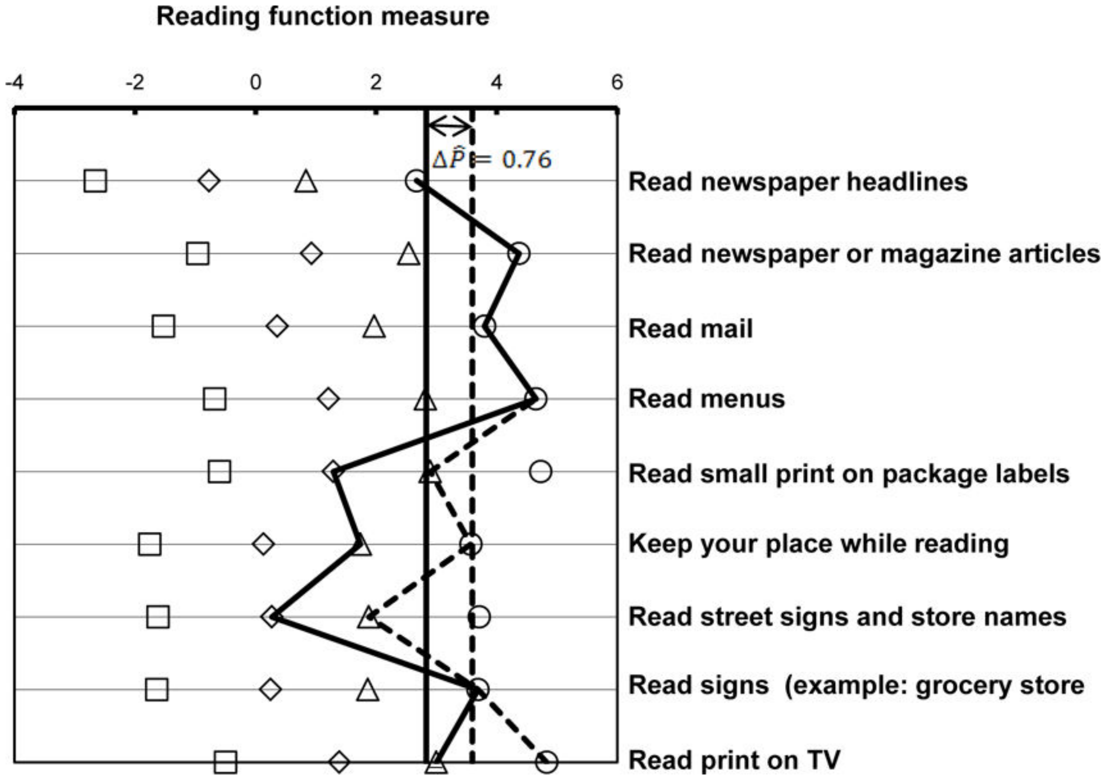


Figure 6.

Keyform plot of a low vision patient's responses to VA LV VFQ reading items at pre-intervention baseline (solid lines) and at post-intervention follow-up (dashed lines). Note that this patient responded "not difficult" (open circles) to five of the items at baseline. The reading function person measure estimated from all nine responses at baseline is the intersection of the solid vertical line with the reading function measure axis. The reading function person measure estimated from all nine responses at post-intervention follow-up is the intersection of the dashed vertical line with the reading function measure axis. The estimated change in the person measure corresponds to the separation of the vertical lines on the reading function measure axis (ΔP = 0.76).

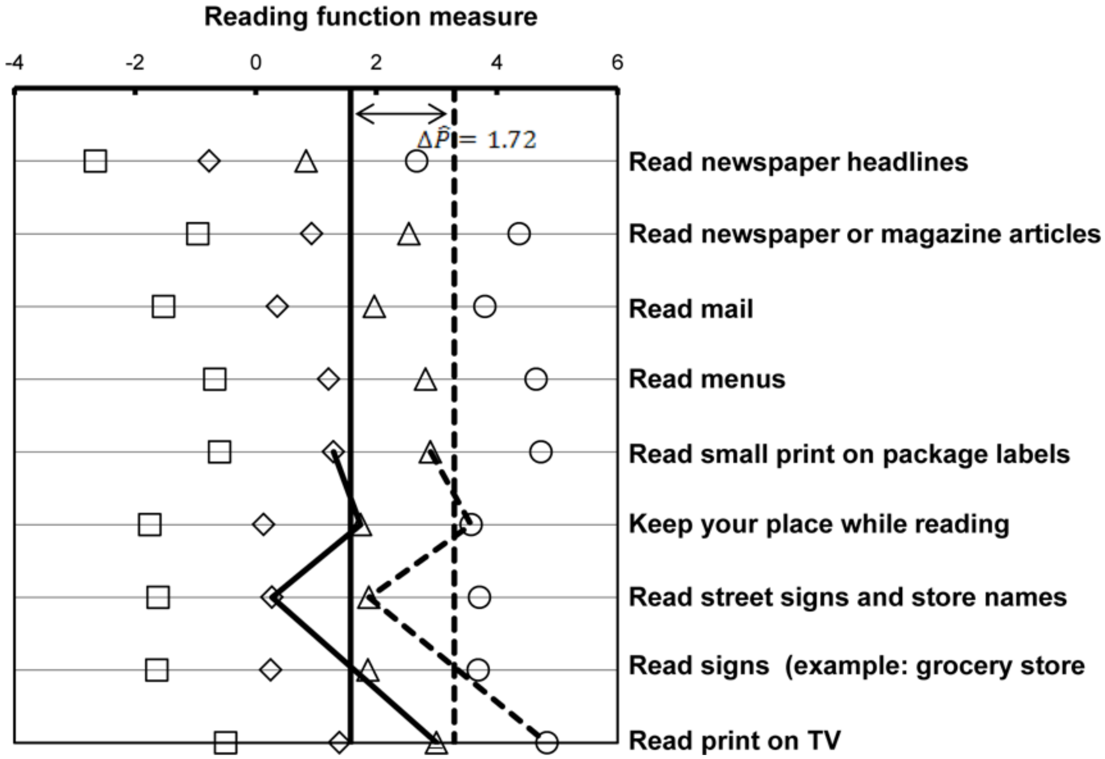


Figure 7.

Keyform plot of the same baseline (solid lines) and follow-up responses (dashed lines) to VA LV VFQ reading items that are plotted in Figure 6, except that the items for which responses at baseline were "not difficult" are removed. The solid vertical line is the average person measure estimate from pre-intervention baseline responses to the four retained items and the dashed vertical line is the average person measure estimate from post-intervention follow-up responses to the same four retained items. By filtering out the items that were rated "not difficult" at baseline, the estimated change in person measure increased to ΔP = 1.72.