Shortening the Leuven Perceptual Organisation Screening Test with Item Response Theory and Confirmatory Factor Analysis

Kathleen Vancleef[1], Nele Demeyere[1], and Luning Sun[2]

[1] Department of Experimental Psychology, University of Oxford

[2] The Psychometrics Centre, University of Cambridge

Author Note

Kathleen Vancleef: https://orcid.org/0000-0002-9943-9341

Nele Demeyere: https://orcid.org/0000-0003-0416-5147

Correspondence concerning this article should be addressed to Luning Sun, The Psychometrics Centre, Cambridge Judge Business School, University of Cambridge, Trumpington Street, Cambridge, CB2 1AG, United Kingdom. Email: ls523@cam.ac.uk

Abstract

The Leuven Perceptual Organisation Screening Test (L-POST) is a reliable and validated test for mid-level visual perceptual deficits after brain injury. However, the test's duration (20-35 minutes) is too long for a screening approach for all patients in clinic practice. Our aim was to shorten L-POST to 10-15 minutes based on statistical criteria of the items. We collected a sample of 3391 participants who completed L-POST. The test consists of 15 subtests with 5 items each. First, we demonstrated high internal consistency of the subtest items through Cronbach's alpha and observed high correlations between scores based on all five items versus a selection of only two items per subtest. This showed that two items per subtest are sufficient. Next, Item Response Theory (IRT) was applied to guide the selection of the items. The highest correlation with full-scale subtest scores was observed when two items were selected following an adaptive testing procedure. A pilot validation in a subsample of participants with low abilities demonstrated adaptive testing has reasonable sensitivity (79%) but limited specificity (54%) in classifying participants with impaired and unimpaired abilities. Last, we reduced the number of subtests through factor analysis. We showed that the subtests using Radial Frequency Patterns as stimuli were redundant and could be combined in one subtest. We conclude that L-POST can be shortened to 26 items (7-14 minutes) and when used with fixed items could prove particularly useful for screening purposes.

Keywords

Short form, IRT, Factor analysis, Classical Test Theory, Perceptual Organisation

Public Significance Statement

Tests for visual problems after brain injury are time-consuming and therefore not often used. In the current research study, we have used sophisticated statistical analyses to shorten an existing test in the best possible way. A shorter test will help remit underdiagnosis and enable access to appropriate treatment.

Visual perception, the processing of visual information in our brain, roughly includes three stages, often labelled low- , mid-, and high level visual perception (Biederman, 1987; Wagemans, Wichmann, & Op de Beeck, 2005). This hierarchical view on visual perception is rooted in our knowledge of visual processing pathways in the brain: signals are transmitted from the retina through the lateral geniculate nucleus and through the early visual cortex (V1-V2) in the occipital pole to higher visual areas (V3-V5) and to other specialised areas in the parietal and temporal lobes (dorsal and ventral pathways) (Kolb & Whishaw, 2003). *Low-level visual perception* or vision refers to processing taking place after light has been converted to electrical signals in the retina. It encompasses coding of simple features like orientation, colour, texture, and size and takes place in the early visual cortex. *Mid-level perception or perceptual organisation* refers to how our brain structures these basic features into coherent wholes (Wagemans, 2015a, 2018). A whole range of visual processes are included in perceptual organisation: perceptual grouping of features belonging to the same object, integration of separate elements into one contour, segregation of two objects based on differences in texture, figure-ground segregation, motion perception, shape perception, etc. In the next stage of *high-level visual perception*, these coherent wholes are linked to existing knowledge of objects or faces for recognition (ventral stream or 'What'-stream) or to our spatial awareness and motor system for localisation and interaction with objects (dorsal stream or "Where"-stream) (Goodale & Milner, 1992; Wagemans, 2015b). In the current manuscript, we focus on assessment of mid-level perception or perceptual organisation.

Assessments for perceptual organisation processes are included in a few neuropsychological batteries. For instance the Visual Object and Space Perception Battery (VOSP, Warrington & James, 1991) starts with a Screening test that assesses intact perceptual grouping and figure-ground segregation. White noise stimuli - similar to static noise on an analog TV - are presented to the respondent. Some stimuli show just white noise, while others display the letter "X" hidden in the white noise. Dot Counting and Incomplete Letters of VOSP also tap into perceptual grouping. In Incomplete Letters, patients have to recognize letters that are degraded by removing square-shaped parts of the letters. In the Dot counting task, patients have to count the number of dots on the page ranging from five to nine dots. The Birmingham Object Recognition Battery (BORB, Riddoch & Humphreys, 1993) includes 13 tasks covering low- to high-level visual perception. The most relevant task for assessing perceptual organisation is Overlapping Figures, in which two or three letters, shapes or drawings of objects are presented adjacent to each other or overlapping. The respondent needs to group line fragments of the same shape, letter or object together to be able to identify the overlapping figures. Other neuropsychological tests that include visual perception subtests are Rivermead Perceptual Assessment Battery (Whiting, Lincoln, Bhavnani, & Cockburn, 1986), Loewenstein Occupational Therapy Cognitive Assessment (Katz, Itzkovich, Averbuch, & Elazar, 1989), Occupational Therapy Adult Perceptual Screening Test (Cooke, McKenna, & Fleming, 2005), and Brain Injury Visual Assessment Battery for Adults (Warren, 1998). However, none of the above tests measures a wide range of perceptual organisation processes with demonstrated psychometric qualities (Torfs, Vancleef, Lafosse, Wagemans, & De-Wit, 2014).

Five years ago, the Leuven Perceptual Organisation Screening Test was developed as a test specific for perceptual organisation with demonstrated reliability and validity (Torfs et al., 2014; Vancleef et al., 2015). In 15 subtests, several processes of perceptual organization are measured (Figure 1A). Stimuli were carefully designed based on theoretical and empirical work in cognitive neuroscience of visual perception. The same matching-to-sample task is used in every subtest:

Participants have to decide which of the three alternative stimuli at the bottom of the screen is most similar to the target stimulus at the top of the screen (Figure 1B). The test is freely available at https://psytests.be/clinicians/ and is described in detail in elsewhere (Torfs et al., 2014). The test has been well-received with over 600 registered clinicians and translations in 7 languages to date.

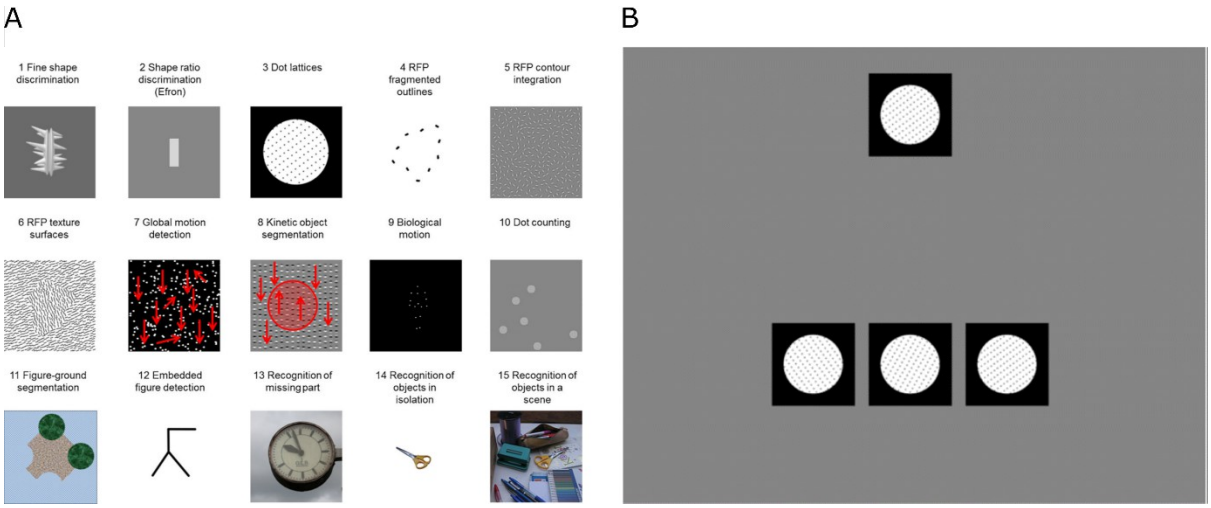
We have previously demonstrated moderate internal consistency of the total L-POST score (Cronbach's alpha = 0.76) in a normative sample of 1,567 healthy volunteers and good test-retest reliability (r = .77) in 20 patients with brain injury (Vancleef et al., 2015). Convergent and discriminant validity was investigated in 58 patients with brain injury. We observed moderate but significant correlations (average r = 0.47) between the total score on L-POST and perceptual organisation subtests of BORB, VOSP and Rey-Complex Figure Test (Anderson, Anderson, & Garth, 2010), while lower and mostly non-significant correlations were observed with measures of spatial attention, executive functions, language, memory, number skills, and praxis (average r = 0.2, only four out of 31 had $p < 0.05$). This indicated that L-POST is specific for visual problems and that scores are not highly influenced by other cognitive impairments.

Although the high number of registered clinicians seems to indicate good adoption of L-POST, 83% of the registered clinicians have only used L-POST three times or less. This suggests that L-POST is not considered suitable for routine use in clinical practice but does not answer the question on which characteristics of L-POST make is unsuitable. We recently conducted qualitative interviews with health care professionals on the barriers and facilitators for mid and high-level visual perception screening after stroke (Vancleef, Colwell, Hewitt, & Demeyere, 2019). Thematic analysis showed that an ideal mid and high-level visual perception assessment should take no longer than 10-15 minutes, require minimal training and equipment, be portable, be suitable for bedside testing, inclusive for participants with aphasia, and evidence-based. Although L-POST only covers mid-level visual perception, the test fits most of the criteria above except from the time requirement. A typical L-POST assessment lasts 20-45 minutes, which is well beyond the ideal testing time of 10-15 minutes. This encouraged us to explore the possibility to shorten L-POST.

In order to reduce the administration time to 10-15 minutes, we would need to reduce the number of items to approximately 30 items or two items per subtest instead of five. Nunnally and Bernstein (1994) provided formulas to estimate the reliability of a hypothetical short form of a test based on reported reliability and validity of the long form (Vancleef et al., 2015). Such estimates are useful as an a priori evaluation of the pay-off between the gained practical efficiency and losses in statistical accuracy (Kruyen, Emons, & Sijtsma, 2013; Smith, McCarthy, & Anderson, 2000) The internal consistency of a 30-item L-POST is estimated at 0.91 compared to 0.96 in the long form, test-retest reliability is estimated at 0.56 compared to 0.76 in the long form, the convergent and discriminant validity is estimated at 0.46 and 0.19 compared to 0.48 and 0.20 in the long form, and the correlation between the short and long form of L-POST is estimated at 0.87. The estimates suggest that a reduction to 30 items and time saving of 50% to 66% only would give a minimal loss of reliability and validity compared to the long form of L-POST (with the exception of test-retest reliability) and seems a worthwhile route to explore.

Figure 1

Leuven Perceptual Organisation Screening Test



Note. (A) Overview of subtests. The red arrows indicate direction of motion of the elements in the image; 1) Fine Shape Discrimination: discriminate novel shapes that differ in fine local aspects of the shape, but with similar global properties; 2) Shape Ratio Discrimination (Efron): judge the ratio of the width and length of a rectangle and select the alternative with the same ratio; 3) Dot Lattices: group dots based on proximity to perceive a general orientation of the pattern and select the alternative with the same general orientation; 4) RFP Fragmented Outlines: group line elements into a global shape using the principle of good continuation; 5) RFP Contour Integration: group collinear Gabor elements to perceive a non-familiar shape and segment it from a background of random oriented Gabor elements; 6) RFP Texture Surfaces: segregate an irregular shape from the background based on texture differences; 7) Global Motion Detection: detect the movement direction of coherently moving dots (with a common fate) in random-dot kinematograms; 8) Kinetic Object Segmentation: group coherently moving Gabor elements inside and outside a kinetic boundary to give rise to the percept of a shape surrounded by a background; 9) Biological Motion: discriminate moving dots making up a walking human from spatially scrambled moving dots; 10) Dot Counting: recognize the number of dots presented during short flashes of 200 milliseconds; 11) Figure-Ground Segmentation: correctly interpret the figure-ground relations to complete the shape that is occluded by circular discs; 12) Embedded Figure Detection: select the complex geometric line pattern in which a simple target pattern is embedded; 13) Recognition of Missing Part: select the alternative in which the same detail is omitted as in the target object; 14) Recognition of Objects in Isolation: recognise everyday objects on a white background. This task serves as a control task for 15) Recognition of Objects in a Scene: recognise the same everyday objects in a natural scene. (B) Example trial of Dot Lattices subtest demonstrating the matching-to-sample principle.

Short forms of psychological tests have long been popular due to a need to increase efficiency of assessments, reduce costs, and minimise burden on patients (Kruyen et al., 2013). Typically, short forms are developed by either reducing the number of subtests, reducing the number of items within a subtest, or both. Development has most often been guided by classic test theory in that short form authors select items or subtests with a high correlation with the overall score of the full test in order to maintain reliability of the test (Ward, Selby, & Clark, 1987). Short form development has also been guided by how representative items or subtests are for the content of the full test. The representative value of an item/subtest can be determined via multiple regression (Jones, 1962), factor analysis (Sellbom & Tellegen, 2019), Item Response Theory (IRT) (Thomas, 2019), or expert judgement of the content of the items/subtests (Kruyen et al., 2013). Last, ad hoc criteria like duration or item/subtest number (e.g. administrating every third subtest, only odd items, only first 10 items etc.) have been used to shorten tests (Kruyen et al., 2013; Satz &

Mogel, 1962; Ward et al., 1987). The sole use of statistical methods make a short form vulnerable to bias of selecting similar items and in that way narrowing the test and reducing construct validity (Kruyen et al., 2013; Smith et al., 2000). Expert judgements on the relevance of items for measurement of the overall construct can avoid this. Most often test constructors decide themselves which items to retain based on perceived relevance (in combination with statistical information, Kruyen et al., 2013). Selection of items based on a combination of statistical information and expert judgements on construct validity has been proposed as the recommended approach (Coste, Guillemin, Pouchot, & Fermanian, 1997; Smith et al., 2000).

Item Response Theory (IRT) has been presented as the future of short form development avoiding many issues with short form development based on Classical Test Theory or factor analysis (Clark & Watson, 2019; Smith et al., 2000). The underlying principle of IRT is that performance on an item reflects a person's ability on an underlying latent variable (Embretson, 1996). This relationship is presented in an Item Characteristic Curve (ICC). An ICC shows that the probability of answering correctly on an item (presented on the y-axis) increases monotonically with increasing ability (often denoted as theta, presented on the x-axis) on the underlying latent variable. The relationship is expected to approximate a logistic function rather than being linear. Different IRT models have been described in the literature (e.g. Rasch model, Partial Credit Model, Graded Response Model, 3-parameter logistic model). The key parameters of an IRT model are item difficulty and item discrimination (Reise & Waller, 2009). Item difficulty is the level of ability at which 50% of the respondents endorses the correct answer. Difficult items require a high level of ability to reach a percentage of 50%, while easy items only require low levels of ability to reach a percentage of 50%. Item discrimination refers to the slope of the curve. Low item discrimination means that the probability of answering correctly on an item increases only slowly with increasing ability, so resulting in a rather flat curve. Such an item does not discriminate well between people with different abilities who will all have similar probabilities of answering the item correctly. High item discrimination means that the probability of answering correctly on an item increases rapidly with increasing ability, so resulting in a rather steep curve. Such an item discriminates well between people with different abilities and is therefore more informative. Compared to Classical Test Theory where each item contributes equally to the overall score (i.e. a sum of the scores of each item), IRT allows items to contribute differently to the overall score depending on the item parameters.

With IRT, we can identify the most informative items in a test and retrain these items in a short form of the test. The most informative items can be chosen for an individual with average ability, but an additional advantage of IRT is that the choice of items can be adapted to match the person's ability. This is called computer-adaptive testing (CAT) (van der Linden & Glas, 2000). For instance, at the start of a test we do not know a person's ability, so presenting a problem of average difficulty will be most informative. If the person solves the problem correctly, it is more likely that this person has above average ability, so the most informative next problem to present would be one of above average difficulty. While for a person solving the first problem incorrectly, a problem of below average difficulty would be more appropriate. Hence, Computer Adaptive Testing (CAT) provides a highly efficient method of assessment. In addition, CAT leads to a better user experience by presenting personalised problems that challenge the individual without them being discouraged by too difficult problems or disinterested by too easy problems. Furthermore, CAT gives more accurate results, reduces fatigue (because tests are shorter), and increases flexibility (Linacre, 2000).

CAT can provide an answer to the problem that only choosing items with high item-total correlations (Classical Test Theory) or high loadings (factor analysis) would reduce the diversity of a test and narrow the content domain being measured. With CAT, no items are permanently removed from the item bank and therefore the diversity and construct validity of the test are retained while still reducing the test length.

The aim of the current research is to develop a short form of L-POST by finding the items that are most informative and reduce duplication of information. As a first step, we investigate internal consistency of each subtest and evaluate whether a sum score of only two items per subtest gives comparable results to a sum score of all five items included in each subtest. This approach follows the principles of Classical Test Theory. Next, we investigate the difference in information value among the items within one subtest and determine whether we can optimize the selection of those two items by selecting the most informative items via IRT and adaptive testing. Finally, we evaluate if items of different subtests can be combined under common factors to reduce the number of subtests a participant would need to complete.

Methods

Sample Characteristics

For this secondary analysis, we considered a convenience sample of 5119 L-POST sessions. These had been completed online by both control and clinical participants between September 2012 and December 2017. The following exclusion criteria were applied in the reported order:

1. Sessions where the participant reported that they had not taken the test seriously (e.g. for demonstration purposes) were excluded (n = 659).
2. Sessions where the participant reported technical problems were excluded (n = 310).
3. Sessions with missing accuracy data were excluded (n = 18). Occasional loss of internet connectivity could have caused an answer to a trial not to be registered.
4. Sessions where the participant reported they had to scroll to see all images on the screen were excluded (n = 141). It was required that participants could see all images simultaneously to avoid that short-term memory deficits were influencing performance on our visual perception task.
5. Sessions where the participant reported issues in the comments' box that might have influenced performance were excluded (n = 52). Examples of comments of excluded sessions were "I didn't bring my reading glasses", "Sometimes issues with clarity due to glare", "First time using a computer, so sometimes made mistakes", "This is a test run", "Accidentally selected wrong option once".
6. Sessions where the participant reported interruptions higher than 2 on a 7-point Likert scale ranging from 'not at all' (level 1) to 'continuously' (level 7) were excluded (n = 352). This reflected moderate to severe levels of interruption.
7. Repeated sessions by the same participant were excluded (n = 214). Only the data from the first session were included in the analysis.

This resulted in 3391 sessions from an equal number of participants. Participants had a median age of 31.05 years old (interquartile range: 21.96 to 49.55 years) and 65% of them identified as women. The median number of years in full-time education was 14 years (interquartile range: 12 to 17 years). Most participants resided in Belgium (47%), Hungary (11%), Turkey (9%), United States (7%), Italy (6%), or United Kingdom (5%). The remaining 15% of participants was spread over 56 countries. 87% of participants reported to be right-handed, 11% left-handed and 2% ambidextrous. 44% reported to have good vision, 54% good vision with glasses or contact lenses, 3% reported impaired vision due to an ophthalmological condition like glaucoma, cataract or macular degeneration. 43 participants (1%) reported having had a brain injury (23 ischemic stroke, 5 traumatic brain injury, 3 anoxia or hypoxia, 1 brain tumour, 1 dementia, 10 other). To maximise the range of scores on L-POST and reduce ceiling effects, participants with visual or neurological conditions were included in our sample.

Instruments

L-POST is a freely-available online test for perceptual organisation. In L-POST, different processes in perceptual organization were measured in 15 subtests (Figure 1A). A short questionnaire at the start asked for biographical and medical details. In the subsequent perceptual organisation screening test all items were presented in a matching-to-sample configuration: a target item centrally at the top of the screen and three alternatives aligned in a row below the target (Figure 1B). To reduce cognitive load, the matching-to-sample task was the same in all subtests, with exactly the same instruction: “Choose the alternative that is most similar to the target stimulus.” We emphasized that no exact matching is necessary through five practice trials. Subtests were administered in a random order, with the restriction that the ‘Recognition of objects in a scene’ subtest always preceded the ‘Recognition of objects in isolation’ subtest. The correct alternative was located in one of the three possible locations. The location was chosen randomly in each item. For each subtest, five items were presented in a block design. L-POST ran in any browser and there were no hardware requirements although a screen that can display all stimuli simultaneously was recommended. The test was available in English, Dutch, Turkish, Hungarian, German, French, Portuguese, and Italian. Participants participated in the L-POST tests unsupervised, with a device, location and time of their choice and in their preferred language. All procedures were approved by Commission for Medical Ethics of [location removed to enable double-blind review] (ML8800).

Data-analyses

Preprocessing

The raw L-POST data contained accuracy information (correct or incorrect) for each of the 75 items of L-POST. The 75 items were grouped in 15 subtests with five items in each subtest (Figure 2). Subtests were categorised under four domains based on a previously reported factor analysis (Vancleef et al., 2015).

Data were downloaded from the L-POST online database and were checked for missing values, inconsistencies and outliers following the principles set out by Jonge and Loo (2013).

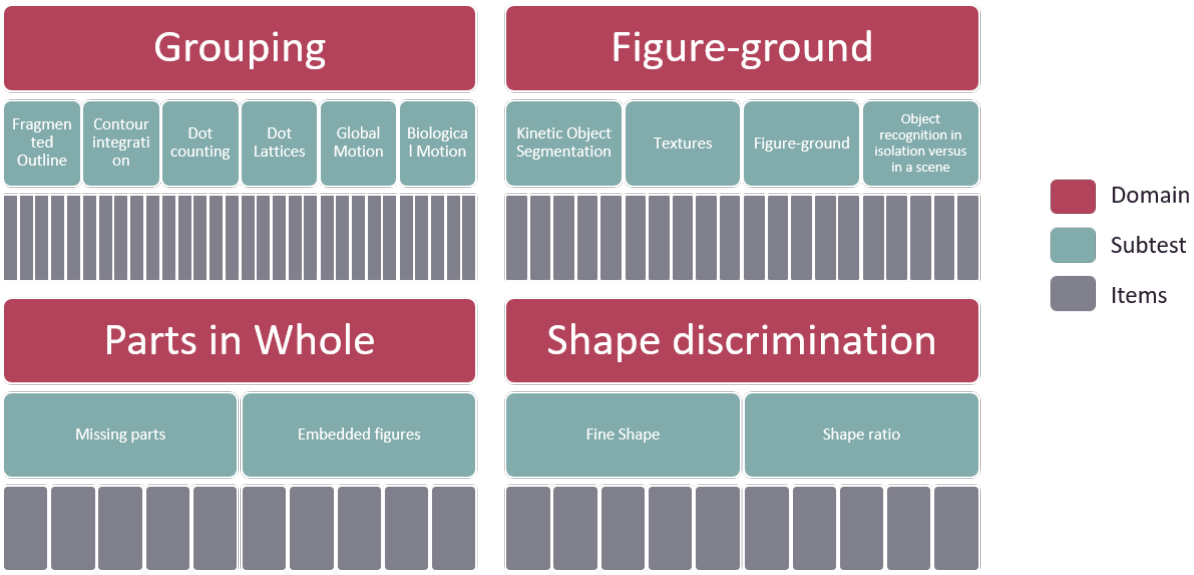
In contrast to other subtests, scores for the Recognition of Objects in Isolation and Recognition of Objects in Scene tasks were not analysed in isolation. Instead the performance on both tasks relative to one another was compared to asses figure ground segregation. For instance, one item in the Objects in Isolation task shows scissors on a white background. In the Object in Scene task the same scissors are shown on a desk. The purpose was not to assess a person’s

performance in each task individually, but to determine whether they would perform better when an object (i.e. the scissors) was presented on a white background compared to a busy background (i.e. the desk). This comparison measures the participant's ability to segregate an object from a background. If a participant gave an incorrect answer for an object in a scene while a correct answer was given for the same object in isolation, they would score 0 (poor performance) on the ability to segregate objects from a background which is noted in this paper as '14–15. Object recognition in a scene' task All other combinations of answers were seen as good segregation ability and received the score of 1. With five objects shown (with and without a background), the scores for the '14–15. Object recognition in a scene' task ranged between 0 (poor performance) and 5 (good performance), just like the other L-POST tasks and can be interpreted in the same way.

Because L-POST was developed to screen for deficits in perceptual organisation after brain injury, the items were designed to be fairly easy for healthy participants (average cut-off scores for L-POST subtests are 4 out of 5 correct). We therefore expect a skewed distribution of scores and also perform our analyses on subsamples of participants with imperfect scores.

Figure 2

Data structure



Internal Consistency and Reduction to Two Items Per Subtest

First, we evaluated the internal consistency of each subtest following Classical Test Theory. We calculated Cronbach's alpha based on tetrachoric correlations. Acceptable levels of alpha ranged between 0.7 and 0.95 (Tavakol & Dennick, 2011). Given acceptable consistency, we subsequently evaluated if reducing the number of items per subtest to two items would be appropriate. We calculated polychoric correlations between the sum scores on each subtest (ranging from 0 to 5) and the sum scores of only two items (ranging from 0 to 2). Polychoric correlations measure the association between two categorical variables with two or more categories. It is based on the assumption that the underlying latent variables follow a bivariate normal distribution. Different

selection methods for the two items were explored: (a) choosing the first two items of each subtest, (b) choosing the first and last item of each subtest to balance potential learning effects in the data, and (c) choosing two items at random. High correlations would indicate that reducing the number of items to two items is appropriate.

IRT Analyses

Next, we evaluated reducing the number of items via Item Response Theory (IRT). The ltm R package was used to estimate the item parameters based on a two-parameter logistic (2PL) model for each subtest. A 2PL model estimated the difficulty and discrimination parameters of an item in a subtest. Subsequently, theta estimates for each person were calculated for different combinations of items. Theta gave an indication of a person's ability on the underlying construct. In our case, this was a person's ability in structuring a visual scene (perceptual organisation). In contrast to sum scores, theta estimates took into account the different psychometric properties of the items. Theta estimates were calculated based on all five items within a subtest, the two items within a subtest that had the highest item discrimination, the most difficult two items within a subtest, the first two items of the subtest, the first and last item of the subtest, two randomly selected items from the subtest, two items selected following an adaptive procedure. The adaptive procedure following maximum fisher information (MFI) first chose the item with the highest information for a person with average ability. Next, a response was simulated and a new theta estimate calculated. The second item was then chosen to be the one with the most information value for the new theta estimate. IRT assumptions were checked through a unidimensional factor analysis for each subtest and by calculating correlations between residuals.

Pilot Validation

The primary aim of L-POST is to screen patients with brain injury for mid-level visual perceptual deficits, so we evaluated sensitivity and specificity of our 2-item subtests selected through IRT in classifying responses as impaired and unimpaired in a subsample of participants of low abilities (with at least one erroneous response in the subtest). As a gold standard we used the classification based on the 5-item version of each subtest. In line with previous recommendations (Vancleef et al., 2015) we set our cut-off at the 10^{th} percentile, meaning a score is considered unimpaired if at least 10% of the observations in the healthy sample were found above that score. Next, we calculated sensitivity and specificity for each subtest and each item selection method (the two items within a subtest that had the highest item discrimination, the most difficult two items within a subtest, the first two items of the subtest, the first and last item of the subtest, two randomly selected items from the subtest, two items selected following an adaptive procedure).

Confirmatory Factor Analyses

Subsequently, we explored reducing the number of subtests via Confirmatory Factor Analysis (CFA). Because our item scores were categorical (0 or 1) model parameters were estimated by diagonally weighted least squares. However, to compute robust standard errors, and a mean- and variance-adjusted test statistic, the full weight matrix was used. Model fit was evaluated and compared on robust estimates of fit indices χ^2, comparative fit index (*CFI*), Tucker–Lewis index (*TLI*), and root mean square error of approximation (*RMSEA*). For cut-off criteria of these fit indices, we followed the guidelines of Hu and Bentler (Hu & Bentler, 1999), who suggested .95 for *CFI* and *TLI* and a cut-off close to .06 for *RMSEA* to conclude a good fit between the hypothesized model and the observed data. As a significant χ^2 relative to the degrees of freedom indicated a difference between

the observed and the model implied variance–covariance matrices, a good fit was associated with a non-significant value. However, because this measure is sensitive to sample size, careful interpretation is advisable. We fitted three factor models (see Results for details on the models). Analyses were performed with lavaan (Rosseel, 2012), a structural equation modelling package for the statistical software R (R Development Core Team, 2011).

Results

Internal Consistency and Reduction to Two Items Per Subtest

Cronbach's alpha indicated good internal consistency (average alpha = 0.81, see Table 1) except for Subtest 2 Shape Ratio Discrimination (alpha = 0.6). The low internal consistency in this subtest resulted from a very low correlation between the last two items in the subtest (Item 18 and Item 19, rho = -0.0064). These items were particularly easy with an average percentage correct of 99%, not one participant solved both items incorrectly, indicating low variance and therefore a low correlation. Polychoric correlations between the sum score based on all five items and the sum score based on two items were high, with most of them above 0.9 (Table 1) irrespectively of how the two items were selected (first two items, first and last item, at random). We calculated the same correlations per subtest based on a subsample with imperfect scores on that subtest (Table 2). The sample sizes ranged from 236 for Kinetic object segmentation to 1405 for Embedded figure detection. Although the average correlation was lower (0.64), we observed the same pattern of results: similar correlations for different selections of items. Taken together, these results indicate that, under the framework of Classical Test Theory, all items within one subtest are very similar and therefore interchangeable to an extent. It shows that two items per subtest are sufficient to reproduce the measurement of perceptual organisation with five items in each subtest, especially in a mixed sample of participants with high and low abilities.

Table 1

Internal Consistency of Subtests.

Subtest	Cronbach's alpha	Correlations between sum score of five items and sum score of		
		First two items	First and last items	Two items selected at random
1. Fine shape discrimination	0.76	0.93	0.94	0.90
2. Shape ratio discrimination (Efron)	0.60	0.95	0.86	0.88
3. Dot lattices	0.79	0.88	0.91	0.91
4. RFP fragmented outline	0.84	0.93	0.93	0.92
5. RFP contour integration	0.82	0.92	0.93	0.92
6. RFP texture surface	0.79	0.92	0.92	0.92
7. Global motion detection	0.94	0.96	0.97	0.96
8. Kinetic object segmentation	0.91	0.95	0.96	0.94
9. Biological motion	0.85	0.95	0.90	0.91
10. Dot counting	0.82	0.88	0.91	0.91
11. Figure-ground segmentation	0.87	0.92	0.89	0.93
12. Embedded figure detection	0.74	0.85	0.85	0.88
13. Recognition of missing part	0.82	0.94	0.92	0.91
14–15. Object recognition in a scene	0.86	0.93	0.95	0.94
Average	0.81	0.92	0.92	0.92

Table 2

Correlations between Sum Score of Five Item and Sum Score of Two Items for Subsample with Low Abilities.

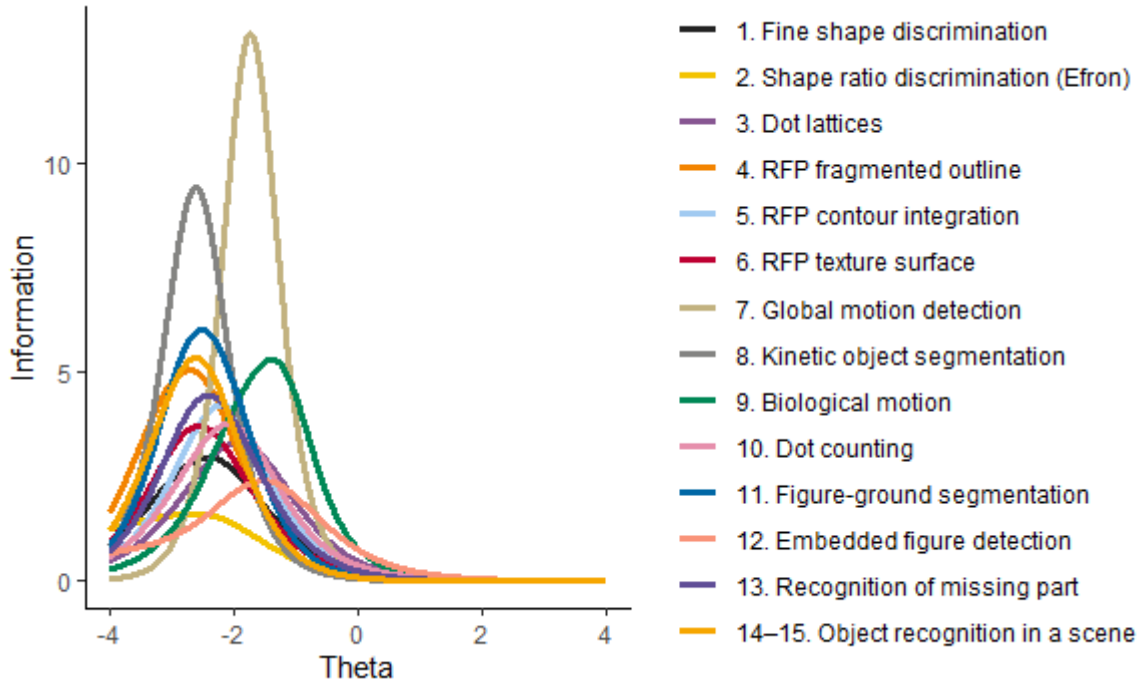
Subtest	N	First two items	First and last items	Two items selected at random
1. Fine shape discrimination	670	0.59	0.70	0.57
2. Shape ratio discrimination (Efron)	527	0.63	0.48	0.39
3. Dot lattices	946	0.57	0.65	0.68
4. RFP fragmented outline	330	0.42	0.69	0.60
5. RFP contour integration	698	0.68	0.65	0.67
6. RFP texture surface	599	0.65	0.68	0.64
7. Global motion detection	572	0.74	0.80	0.83
8. Kinetic object segmentation	236	0.79	0.64	0.70
9. Biological motion	1278	0.71	0.47	0.72
10. Dot counting	892	0.67	0.76	0.69
11. Figure-ground segmentation	380	0.51	0.72	0.64
12. Embedded figure detection	1405	0.59	0.58	0.65
13. Recognition of missing part	609	0.60	0.60	0.61
14–15. Object recognition in a scene	302	0.64	0.68	0.66
Average		0.63	0.65	0.65

IRT and Adaptive Testing

The 2PL IRT models we fitted to each subtest all showed good fit with our data and assumptions were met (all correlations between residuals < 0.2, and goodness-of-fit indices of unidimensional factor analyses per subtest were good, with ≥ 0.97 for CFI, ≥ 0.95 for TLI, and ≤ 0.4 for RMSEA). Items had varying difficulties but were all fairly easy (median = -2.4, *IQR* = [-2.79, -2.15]) with one outlier of -6.29 (Item 19) indicating this item is extremely easy, as we already showed above. All discrimination parameters were positive (median = 1.94, *IQR* = [1.50, 2.35]) as expected, indicating that people with higher levels of ability are more likely to solve the item correctly (monotonocity). Item 19 had very little discrimination power (0.69) as can be expected from an easy item that nearly all participants solved correctly (see Table S1 for all parameter estimates). The test information curves in Figure 3 illustrate that subtests are most informative for participants with low ability levels. This reflects the intended purpose of L-POST as a screening instruments for impaired visual perception in patients with brain injury rather than as an instrument to measure subtle individual differences in people with no impairments in visual perception.

Figure 3

Test Information Functions for L-POST Subtests.



Note. These test information curves show the amount of information that is provided by each subtest for people with different levels of ability (theta). Theta does not refer to a general ability, but refers to a specific perceptual organisation ability depending on which subtest you are considering. For instance, when interpreting the test information function of '3 Dot lattices', theta refers to perceptual grouping ability, while when considering the test information function of '7 Global motion detection', theta refers to motion perception ability. All subtests provide most information for people with low levels of ability (negative theta) rather than for people with high levels of ability (positive theta). Subtest 7 'Global motion detection' is the most informative subtest. Subtest 2 'Shape ratio discrimination (Efron)' is the least informative subtest.

Correlations between theta estimates (representing a person's ability) based on all five items and based on different selections of two items are presented in Table 3. Comparison between the different selections of items shows that on average correlations are similar between randomly selected items, selection of the two most discriminating items, and selection of the two most difficult items, indicating items are similar. However, the highest correlation was observed when items are selected based on adaptive testing. Adaptive testing allows optimal selection of the items that cater to the ability of each participant. This even has beneficial effects when only one item (the second item) is adaptively selected (as the first item in adaptive testing was the same across all participants). In our subsamples of participants with low abilities, we see the highest correlations when the two most discriminating items are selected (0.75), rather than adaptively selected items (0.46) (Table 4). One possible reason is that the first item in adaptive testing is always the one that has the highest information for a person with average ability. With only two items allowed per participant, adaptive testing was less informative in subsamples with poor abilities.

Table 3

Correlations Between Theta Estimate of All Five Items of a Subtest and Theta Estimate of Two Items

Subtest	First two items	First and last item	Two items selected at random	Two most discriminating items	Two most difficult items	Adaptive testing with two items
1. Fine shape discrimination	0.75	0.84	0.71	0.77	0.76	0.84
2. Shape ratio discrimination (Efron)	0.71	0.50	0.68	0.79	0.79	0.90
3. Dot lattices	0.65	0.73	0.75	0.81	0.81	0.81
4. RFP fragmented outline	0.72	0.77	0.70	0.81	0.80	0.90
5. RFP contour integration	0.74	0.76	0.74	0.74	0.73	0.78
6. RFP texture surface	0.73	0.75	0.71	0.70	0.69	0.80
7. Global motion detection	0.84	0.86	0.82	0.81	0.81	0.84
8. Kinetic object segmentation	0.74	0.83	0.73	0.64	0.63	0.84
9. Biological motion	0.86	0.71	0.80	0.83	0.81	0.90
10. Dot counting	0.64	0.71	0.76	0.78	0.75	0.84
11. Figure-ground segmentation	0.68	0.54	0.79	0.73	0.73	0.85
12. Embedded figure detection	0.66	0.67	0.75	0.83	0.83	0.84
13. Recognition of missing part	0.80	0.71	0.77	0.62	0.61	0.80
14–15. Object recognition in a scene	0.70	0.76	0.76	0.79	0.79	0.83
Average	0.73	0.72	0.75	0.76	0.75	0.84

Table 4

Correlations Between Theta Estimate of All Five Items of a Subtest and Theta Estimate of Two Items in a Subsample with Low Abilities

Subtest	N	First two items	First and last item	Two items selected at random	Two most discriminating items	Two most difficult items	Adaptive testing with two items
1. Fine shape discrimination	670	0.20	0.62	0.48	0.72	0.72	0.62
2. Shape ratio discrimination (Efron)	527	0.01	0.06	0.37	0.82	0.82	0.71
3. Dot lattices	946	0.23	0.42	0.53	0.68	0.67	0.64
4. RFP fragmented outline	330	-0.07	0.69	0.34	0.78	0.80	0.35
5. RFP contour integration	698	0.54	0.39	0.51	0.76	0.74	0.34
6. RFP texture surface	599	0.51	0.60	0.39	0.78	0.77	0.44
7. Global motion detection	572	0.39	0.49	0.62	0.74	0.75	0.42
8. Kinetic object segmentation	236	0.66	0.08	0.61	0.78	0.78	0.15
9. Biological motion	1278	0.51	0.07	0.68	0.84	0.83	0.64
10. Dot counting	892	0.51	0.65	0.53	0.80	0.76	0.51
11. Figure-ground segmentation	380	0.11	0.59	0.52	0.69	0.69	0.28
12. Embedded figure detection	1405	0.40	0.44	0.55	0.74	0.73	0.62
13. Recognition of missing part	609	0.24	0.30	0.55	0.68	0.68	0.24
14–15. Object recognition in a scene	302	0.40	0.26	0.47	0.66	0.66	0.54
Average		0.33	0.40	0.51	0.75	0.74	0.46

Pilot Validation

We tested the above prediction on classification performance of L-POST in a pilot validation study using the subsample with low abilities. A cut-off at the 10th percentile of theta (-1.28) resulted in very low sensitivity of any 2-item version (average sensitivity over subtests and selection methods = 47%) with on average only 4% of participants being classified as impaired compared to 6.6% based on all 5 items. With fewer items, there was a reduced chance to pick up erroneous responses, hence more people solved all items within a subtest correctly and were classified as unimpaired. Furthermore, with fewer items, the Standard Error of Measurement increased (average 0.93) compared to the long version with 5 items (average 0.86), reducing the confidence in the estimated

theta scores. Therefore, we opted to increase our cut-off to the 20th percentile of theta (-0.84). With this cut-off the sensitivity increases substantially to 69% (averaged over all subtest and item selection methods) with the highest sensitivity of 79.2% for items selected through adaptive testing (Table 5). It is also noted that the specificity for items selected through adaptive testing was only 53.66% in the subsample with low abilities (Table 6), which could potentially explain the high sensitivity observed here. Specificity is on average 66% with the highest specificity when the two most difficult items are selected (85%). Although for a screening test like L-POST, high sensitivity is usually preferred over high specificity to avoid the potential positive cases are missed, caution should be taken to interpret these results, particularly when comparing adaptive testing to the selection of the two most difficult items.

Table 5

Sensitivity of Theta Estimate of Two Items

Subtest	First two items	First and last item	Two items selected at random	Two most discriminating items	Two most difficult items	Adaptive testing with two items
1. Fine shape discrimination	67.46%	91.72%	71.01%	78.11%	78.11%	91.72%
2. Shape ratio discrimination (Efron)	91.25%	50.00%	63.75%	78.75%	73.75%	73.75%
3. Dot lattices	54.01%	73.42%	81.86%	95.78%	95.78%	90.72%
4. RFP fragmented outline	37.23%	78.19%	47.87%	84.04%	84.04%	76.06%
5. RFP contour integration	79.25%	80.66%	76.89%	74.06%	74.06%	86.32%
6. RFP texture surface	70.06%	68.79%	71.97%	66.24%	66.24%	79.62%
7. Global motion detection	88.51%	91.95%	82.38%	77.01%	77.01%	88.51%
8. Kinetic object segmentation	41.95%	76.27%	42.80%	24.58%	24.58%	76.27%
9. Biological motion	84.29%	27.24%	82.05%	98.08%	39.10%	93.91%
10. Dot counting	53.62%	74.89%	64.26%	87.23%	49.79%	62.13%
11. Figure-ground segmentation	32.22%	26.67%	61.48%	55.56%	55.56%	66.30%
12. Embedded figure detection	85.45%	86.36%	70.00%	77.73%	61.36%	61.36%
13. Recognition of missing part	85.16%	77.47%	81.32%	50.55%	50.55%	85.16%
14–15. Object recognition in a scene	31.69%	48.97%	57.61%	62.55%	62.55%	76.95%
Average	64.44%	68.04%	68.23%	72.16%	63.75%	79.20%

Table 6

Specificity of Theta Estimate of Two Items

Subtest	First two items	First and last item	Two items selected at random	Two most discriminating items	Two most difficult items	Adaptive testing with two items
1. Fine shape discrimination	68.06%	48.10%	68.66%	70.86%	70.86%	48.10%
2. Shape ratio discrimination (Efron)	38.93%	82.77%	71.36%	69.57%	73.83%	73.83%
3. Dot lattices	49.22%	50.78%	57.69%	57.97%	57.97%	53.17%
4. RFP fragmented outline	0.00%	100.00%	53.52%	100.00%	100.00%	0.00%
5. RFP contour integration	64.61%	51.44%	63.37%	79.63%	79.63%	44.86%
6. RFP texture surface	59.05%	62.90%	57.69%	87.10%	87.10%	40.95%
7. Global motion detection	42.77%	43.73%	60.13%	69.13%	69.13%	42.77%
8. Kinetic object segmentation						
9. Biological motion	81.57%	92.65%	74.33%	72.26%	96.38%	75.88%
10. Dot counting	75.34%	77.78%	78.69%	79.15%	89.65%	89.50%
11. Figure-ground segmentation	0.00%	100.00%	53.64%	100.00%	100.00%	0.00%
12. Embedded figure detection	62.03%	63.38%	80.34%	78.90%	95.27%	95.27%
13. Recognition of missing part	33.26%	56.67%	59.25%	88.99%	88.99%	33.26%
14–15. Object recognition in a scene	0.00%	0.00%	69.49%	100.00%	100.00%	100.00%
Average	44.22%	63.86%	65.24%	81.04%	85.29%	53.66%

Factor Structure

The first model tested, the domain model, represented the previously reported four perceptual organisation domains: Perceptual Grouping, Figure-Ground Segmentation, Parts in Whole and Shape Discrimination (see Figure 2) (Vancleef et al., 2015). In contrast to the previous report, individual items with scores of 0 or 1 were used instead of subtests with scores ranging from 0 to 5. Fit indices for the model indicated a poor fit of the model to the data (Table 7), suggesting that the current grouping of subtests into four domains is not ideal, though standard loadings were acceptable (Mdn = 0.6; range = [0.34, 0.87]).

Second, we tested a model with subtests as factors. Fit indices indicated a good model fit, as did the standardised factor loadings (Mdn = 0.71, range = [0.47, 0.96]). All inter-factor correlations were within the normal range, except a high correlation (r = 1) between Subtest 6 RFP Texture Surfaces and Subtest 4 RFP Fragmented Outlines, indicating both subtests measure the same construct.

Third, we tested a model with subtest as factors but we combined all subtests using Radial Frequency Patterns (RFP, Subtests 4-6) as stimuli into one factor. Fit indices again indicated a good model fit, as did the standardised factor loadings (Mdn = 0.7, range = [0.47, 0.96], Table S2). All correlations between factors were between 0.23 and 0.81 (Table S3). We accepted this model as our final model.

Table 7

Goodness-of-Fit Indicators of CFA Models

Model	χ^2	*Df*	χ^2/df	*CFI*	*TLI*	*RMSEA*
Domain model	4761.98**	2339	2.04	.83	.83	.017
Subtest model (one factor per subtest)	2429.85*	2254	1.08	.99	.99	.005
Subtest model (Subtests 4, 5, and 6 combined)	2464.35*	2279	1.08	.99	.99	.005

Note.

**p<.001, *p<.05.

CFI = Comparative Fit Index, TLI = Tucker Lewis Index, RMSEA = Root Mean Square Error of Approximation

Discussion

To sum up, the results of our internal consistency analyses indicate that items within subtests are similar and that two instead of five items per subtest are sufficient under the Classical Test Theory. Item Response Theory analysis shows that computerized adaptive testing (CAT) of two items yields better results than selecting the first two times, the first and last items, the two with the highest discrimination, the two most difficult items, or two randomly selected items. In a pilot validation, we demonstrated that a short version of L-POST (2 items per subtest) has reasonable sensitivity in classifying participants with impaired and unimpaired perceptual organisation abilities. A larger item bank with a wider range of difficulties is recommended for measuring more subtle inter-individual differences. Confirmatory Factor Analysis (CFA) suggests that all RFP subtests measure a common factor and are therefore redundant. Taken together, the results suggest that L-POST can best be shortened to 13 subtests (by merging Subtests 4-6) with 2 items for each subtest, and items for an individual should be chosen using CAT. However, an item bank of only five item per factor (15 for the RFP factor) is insufficient for multidimensional CAT and more stimuli will need to be developed for each subtest to achieve this.

The suggested changes will reduce the number of items from 75 to 26 and shorten the estimated time to complete the test to 7-14 minutes. This brings the short form of L-POST in the range of preferred testing time of 10-15 minutes (Vancleef et al., 2019). Psychometric properties of any short form of L-POST will need to be evaluated, before it can be of clinical value in the diagnosis of perceptual organisation deficits after brain injury (Kleka & Paluchowski, 2017; Kruyen et al., 2013; Smith et al., 2000). Estimates of reliability and validity as reported in the Introduction are often overestimating empirical values and should only be used to guide development (Nunnally & Bernstein, 1994; Smith et al., 2000). A psychometric investigation should evaluate internal consistency, test-retest reliability, convergent and discriminant validity, correlation between the short and long form of L-POST, confirmatory factor analysis, all in an independent sample of participants who completed the short form of L-POST (Kruyen et al., 2013; Smith et al., 2000).

When comparing our results to previously published psychometric reports, we noticed that the Cronbach's alpha coefficients reported in Table 1 were considerably higher than the ones reported in our earlier paper (Vancleef et al., 2015) where the average Cronbrach's alpha for the subtests was only .48 compared to .81 in the current sample. This might be surprising given that most short forms of psychological tests have a lower reliability than their full counterpart (Kruyen et al., 2013). However, the reason for this difference can be found in the sample size (1567 in the previous sample and 3391 in the current sample). Simulations have shown that in skewed distributions with positive kurtosis, with five items, and with high internal consistency, the estimated Cronbach's alpha increases substantially with increasing sample size (Sheng & Sheng, 2012). Low sample sizes lead to an underestimation of the real internal consistency. Sheng and Sheng showed that increased sample size helps to offset non-normality. Our current average estimate of .81 is therefore more likely to be correct than our previously reported average estimate of .48. We observed slightly higher reliability than was observed in a review of previously published short forms of psychological tests (average Cronbach's alpha = .78, N = 291, Kruyen et al., 2013).

The final factor solution for L-POST differs from the factor solution in a previous report (Vancleef et al., 2015). We tested the previously reported domain model but achieved only

moderate model fitting (see Table 3). A major difference in our current and past approach lies in the raw data that entered the factor analysis. Previously, CFA was performed on the 15 subset sum scores (range: 0-5), while now we used 80 item scores (75 original items and 5 items that are a comparison between object recognition in isolation versus in a scene; range: 0-1). Using item scores rather than sum scores allowed us to examine relationships among individual items within and between subtests. This additional flexibility resulted in better model fits.

A common criticism of short form development based on statistical criteria is that the construct validity of the test can be inflated. Only choosing items with high item-total correlations (Classical Test Theory) or high loadings (factor analysis) can reduce the diversity of a test and narrow the content domain being measured. It is unlikely that this has been problematic in our study because we demonstrated that all items within one subtest were highly similar and interchangeable (Table 1).

However, we can identify some limitations of our study and provide directions for future research. First, we did not explore any differential effect in different demographic groups via Differential Item Functioning (DIF). It was previously demonstrated in a smaller sample (N = 1567) that the effects of age, gender, testing condition (lab versus online), and education levels only explain between 2% and 6% of the variance of the L-POST scores (Vancleef et al., 2015). We therefore chose not to explore DIF for these variables. The only variable having a substantial effect on L-POST scores was brain injury. Unfortunately, of our sample of 3391 participants only 43 had a brain injury, which is insufficient for exploring differential effects through a DIF analysis. Additional data on L-POST in a larger sample of people with brain injury will allow us to investigate this crucial DIF for brain injury in the future. Second, nearly all L-POST items have a highly skewed distribution and ceiling effects have likely influenced our analyses (Clark & Watson, 2019). For instance, skewed distributions might have inflated the correlations and have given rise to an artefactual difficulty factor (Sellbom & Tellegen, 2019). We therefore have run the same analyses on a subsample of participants with imperfect scores. We showed that it is advisable to develop a larger item bank per subtest with a wider range of difficulties to detect subtle inter-individual difference in the general population and to classify impaired and unimpaired perceptual organisation function. Third, certain methodical choices can be criticized. For instance, item 19 has proven problematic in the correlation analysis and the IRT analysis. One could argue that this item should be removed from a shortened version of L-POST. Also, we opted for a 2PL instead of a 3PL model with a guessing parameter. A comparison between model fits of 2PL and 3PL models showed better fits (lower AIC, lower BIC, and insignificant Chi^2 test for model comparisons) for 2PL models. Furthermore, a selection of 2 items for each of the 15 subtests gave acceptable results. However, there is potential to improve a short version by selecting more items per subtest to increase reliability per subtest, but only include subtests that are likely impaired in patients with brain injury rather than including all subtests. Last, for clinical utility one test which covers the full range of low-, mid-, and high-level visual perception functions would be more relevant. L-POST should therefore be used in conjunction with other tests if a clinician wants to achieve a comprehensive overview of visual and visual perception abilities.

To our knowledge, we are the first to shorten a visual perception test like the ones listed above (e.g. BORB, VOSP, Rivermead Perceptual Assessment Battery, Loewenstein Occupational Therapy Cognitive Assessment, Occupational Therapy Adult Perceptual Screening Test, and Brain Injury Visual Assessment Battery for Adults) through Confirmatory Factor Analysis or Item Response

Theory . Item Response Theory has been used to explore differential response patterns in visual perception tests for different groups of subject (e.g. older and younger adults) via Differential Item Functioning in the Cambridge Face Memory Test and the Vanderbilt Expertise Test for Cars (Cho et al., 2015; Lee, Cho, McGugin, Van Gulick, & Gauthier, 2015). Confirmatory Factor Analysis has also been applied to visual perception assessments to confirm theoretical foundations of the assessment (Klein, Sollereder, & Gierl, 2002; Rapport, Millis, & Bonello, 1998), but not to shorten an assessment. Although, Computer Adaptive Testing is commonly used in educational assessments (e.g. language tests) and several open-source platforms have been made available (e.g. Concerto, University of Cambridge Psychometrics Centre, 2020 and TAO, Open Assessment Technologies, 2020), adoption in neuropsychological practice has been very limited (Thomas, 2019).

Our previous research has shown that the main barriers to adoption of mid- and high level visual perception screening are: lack of time, lack of training, environmental factors (e.g. distractions at bed-side), stroke survivor factors (e.g. aphasia or fatigue), lack of continuation of care and test characteristics (availability of materials, evidence-based, meaningful for daily life, included in clinical guidelines) (Vancleef, Colwell, Hewitt, & Demeyere, 2020). A shortened version of L-POST has the potential to resolve at least a few of these barriers because such a test would be quick (lack of time), easy to administer without any formal training (lack of training), aphasia- and neglect friendly (stroke survivors factors), available online and evidence-based (test characteristics). We therefore believe that a shortened L-POST has the potential to benefit clinical practice and have high adoption rates. For a complete overview of visual and visual perception functions after brain injury, L-POST would need to be complemented by tests for low- and high-level visual perception.

We conclude that L-POST can be shortened to 26 items which corresponds to an administration time of approximately 10 minutes. Adopting a Computer Adaptive Testing procedure yields the highest precision. Future research should focus on developing an item bank with a larger variety of difficulties.

Acknowledgments

KV and ND are supported by the Stroke Association [grant numbers TSA PDF 2017/03, TSA LECT 2015/02]. We would like to thank Rudy Dekeerschieter, Christophe Bossens, and Johan Wagemans for programming support and continued discussions about the future of L-POST.

References

Anderson, P., Anderson, V., & Garth, J. (2010). Assessment and Development of Organizational Ability : The Rey Complex Figure Organizational Strategy Score (RCF-OSS)* Assessment and Development of Organizational Ability : The Rey Complex Figure Organizational Strategy Score (RCF-OSS)*, (May 2013), 37–41.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*(2), 115–117. https://doi.org/10.1037//0033-295X.94.2.115

Cho, S.-J., Wilmer, J., Herzmann, G., McGugin, R. W., Fiset, D., Van Gulick, A. E., … Gauthier, I. (2015). Item response theory analyses of the Cambridge Face Memory Test (CFMT). *Psychological Assessment*, *27*(2), 552–566. https://doi.org/10.1037/pas0000068

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, *31*(12), 1412–1427. https://doi.org/10.1037/pas0000626

Cooke, D. M., McKenna, K., & Fleming, J. (2005). Development of a standardized occupational therapy screening tool for visual perception in adults. *Scandinavian Journal of Occupational Therapy*, *12*(2), 59–71. https://doi.org/10.1080/11038120410020683-1

Coste, J., Guillemin, F., Pouchot, J., & Fermanian, J. (1997). Methodological approaches to shortening composite measurement scales. *Journal of Clinical Epidemiology*, *50*(3), 247–252. https://doi.org/10.1016/S0895-4356(96)00363-0

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, *8*(4), 341–349. https://doi.org/10.1037/1040-3590.8.4.341

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, *15*(1), 20–25. https://doi.org/10.1016/0166-2236(92)90344-8

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Jones, R. L. (1962). Analytically developed short forms of the WAIS. *Journal of Consulting Psychology*, *26*(3), 289–289. https://doi.org/10.1037/h0047992

Jonge, E. de, & Loo, M. van der. (2013). *An introduction to data cleaning with R*. Statistics Netherlands.

Katz, N., Itzkovich, M., Averbuch, S., & Elazar, B. (1989). Loewenstein Occupational Therapy Cognitive Assessment (LOTCA) Battery for Brain-Injured Patients: Reliability and Validity. *American Journal of Occupational Therapy*, *43*(3), 184–192. https://doi.org/10.5014/ajot.43.3.184

Klein, S., Sollereder, P., & Gierl, M. (2002). Examining the Factor Structure and Psychometric Properties of the Test of Visual-Perceptual Skills. *OTJR: Occupation, Participation and Health*, *22*(1), 16–24. https://doi.org/10.1177/153944920202200103

Kleka, P., & Paluchowski, W. J. (2017). Shortening of psychological tests - Assumptions, methods and

doubts. *Polish Psychological Bulletin*, *48*(4), 516–522. https://doi.org/10.1515/ppb-2017-0058

Kolb, B., & Whishaw, I. Q. (2003). *Fundamentals of human neuropsychology* (7th editio). New York, NY: Worth Publishers.

Kruyen, P. M., Emons, W. H. M., & Sijtsma, K. (2013). On the Shortcomings of Shortened Tests: A Literature Review. *International Journal of Testing*, *13*(3), 223–248. https://doi.org/10.1080/15305058.2012.703734

Lee, W. Y., Cho, S. J., McGugin, R. W., Van Gulick, A. B., & Gauthier, I. (2015). Differential item functioning analysis of the Vanderbilt Expertise Test for cars. *Journal of Vision*, *15*(13), 1–19. https://doi.org/10.1167/15.13.23

Linacre, J. M. (2000). Computer-Adaptive Testing: A Methodology Whose Time Has Come. In S. Chae, E. Jeon, & J. M. Linacre (Eds.), *Development of Computerized Middle School Achievement Test*. Seoul, South Korea: Komesa Pres. Retrieved from https://www.cehd.umn.edu/EdPsych/C-Bas-R/Docs/Linacre2000_CAT.pdf

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New-York: McGraw-Hill.

Open Assessment Technologies. (2020). TAO. Retrieved from https://www.taotesting.com/

Rapport, L. J., Millis, S. R., & Bonello, P. J. (1998). Validation of the Warrington theory of visual processing and the visual object and space perception battery. *Journal of Clinical and Experimental Neuropsychology*, *20*(2), 211–220. https://doi.org/10.1076/jcen.20.2.211.1169

Reise, S. P., & Waller, N. G. (2009). Item Response Theory and Clinical Measurement. *Annual Review of Clinical Psychology*, *5*(1), 27–48. https://doi.org/10.1146/annurev.clinpsy.032408.153553

Riddoch, M. J., & Humphreys, G. W. (1993). *Birmingham Object Recognition Battery*. London, United Kingdom: Psychology Press.

Satz, P., & Mogel, S. (1962). An abbreviation of the wais for clinical use. *Journal of Clinical Psychology*, *18*(1), 77–79. https://doi.org/10.1002/1097-4679(196201)18:1<77::AID-JCLP2270180124>3.0.CO;2-R

Sellbom, M., & Tellegen, A. (2019). Factor analysis in psychological assessment research: Common pitfalls and recommendations. *Psychological Assessment*, *31*(12), 1428–1441. https://doi.org/10.1037/pas0000623

Sheng, Y., & Sheng, Z. (2012). Is coefficient alpha robust to non-normal data? *Frontiers in Psychology*, *3*(FEB), 1–13. https://doi.org/10.3389/fpsyg.2012.00034

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment*, *12*(1), 102–111. https://doi.org/10.1037/1040-3590.12.1.102

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Thomas, M. L. (2019). Advances in applications of item response theory to clinical assessment. *Psychological Assessment*, *31*(12), 1442–1455. https://doi.org/10.1037/pas0000597

Torfs, K., Vancleef, K., Lafosse, C., Wagemans, J., & De-Wit, L. (2014). The Leuven Perceptual

Organization Screening Test (L-POST), an online test to assess mid-level visual perception. *Behavior Research Methods*, *46*(2), 472–487. https://doi.org/10.3758/s13428-013-0382-6

University of Cambridge Psychometrics Centre. (2020). Concerto. Retrieved from https://concertoplatform.com/about

van der Linden, W. J., & Glas, C. A. W. (2000). *Computerized Adaptive Testing: Theory and Practice*. (W. J. van der Linden & C. A. W. Glas, Eds.). Monterey, CA: Kluwer Academic Publishers.

Vancleef, K., Acke, E., Torfs, K., Demeyere, N., Lafosse, C., Humphreys, G., … De-Wit, L. (2015). Reliability and validity of the Leuven Perceptual Organization Screening Test (L-POST). *Journal of Neuropsychology*, *9*(2), 271–298. https://doi.org/10.1111/jnp.12050

Vancleef, K., Colwell, M., Hewitt, O., & Demeyere, N. (2019). Current practice and challenges in screening for visual perception deficits after stroke: a qualitative study. *MedRXiv*. https://doi.org/10.1101/19013243

Vancleef, K., Colwell, M. J., Hewitt, O., & Demeyere, N. (2020). Current practice and challenges in screening for visual perception deficits after stroke: a qualitative study. *Disability and Rehabilitation*. https://doi.org/10.1080/09638288.2020.1824245

Wagemans, J. (Ed.). (2015a). *The Oxford Handbook of Perceptual Organization*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199686858.001.0001

Wagemans, J. (2015b). Vision, High-Level Theory of. *International Encyclopedia of the Social & Behavioral Sciences: Second Edition*, (1987), 153–157. https://doi.org/10.1016/B978-0-08-097086-8.43099-0

Wagemans, J. (2018). Perceptual Organization. In *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience* (Vol. 2, pp. 1–70). Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9781119170174.epcn218

Wagemans, J., Wichmann, F. A., & Op de Beeck, H. P. (2005). Visual perception I: Basic principles. In K. Lamberts & R. Goldstone (Eds.), *Handbook of cognition* (pp. 3–47). London, United Kingdom: Sage Publications.

Ward, L. C., Selby, R. B., & Clark, B. L. (1987). Subtest administration times and short forms of the Wechsler Adult Intelligence Scale-Revised. *Journal of Clinical Psychology*, *43*(2), 276–278. https://doi.org/10.1002/1097-4679(198703)43:2<276::AID-JCLP2270430219>3.0.CO;2-U

Warren, M. (1998). *Brain Injury Visual Assessment Battery for Adults*. Lenexa, KS: visAbilities Rehab Services.

Warrington, E. K., & James, M. (1991). *The Visual Object and Space Perception Battery*. Bury St. Edmunds, United Kingdom: Thames Valley Test Company.

Whiting, S., Lincoln, N. B., Bhavnani, G., & Cockburn, J. (1986). Rivermead Perceptual Assessment Battery. *Occupational Therapy In Health Care*, *3*(3–4), 209–210. https://doi.org/10.1080/J003v03n03_18

Appendices

Table S1

Parameter estimates of IRT models per subtest.

Subtest	Item	Difficulty	Discrimination
1. Fine shape discrimination	Item 0	-2.15	1.51
	Item 1	-3.03	1.03
	Item 2	-3.01	1.45
	Item 3	-2.50	1.96
	Item 4	-2.22	1.79
2. Shape ratio discrimination (Efron)	Item 15	-3.38	1.16
	Item 16	-2.88	1.03
	Item 17	-2.26	1.85
	Item 18	-3.83	1.49
	Item 19	-6.28	0.69
3. Dot lattices	Item 5	-2.24	1.27
	Item 6	-2.34	1.40
	Item 7	-1.95	1.92
	Item 8	-1.70	1.82
	Item 9	-1.96	1.76
4. RFP fragmented outline	Item 35	-3.15	2.15
	Item 36	-2.47	1.48
	Item 37	-3.25	2.21
	Item 38	-2.64	2.57
	Item 39	-2.23	2.40
5. RFP contour integration	Item 10	-2.23	1.58
	Item 11	-2.26	2.03
	Item 12	-2.19	2.35
	Item 13	-2.19	1.46
	Item 14	-2.24	1.61
6. RFP texture surface	Item 30	-2.35	2.13
	Item 31	-2.87	1.17
	Item 32	-2.76	2.33
	Item 33	-2.25	1.51
	Item 34	-2.66	1.37
7. Global motion detection	Item 50	-1.73	2.22
	Item 51	-1.72	3.06
	Item 52	-1.79	3.50
	Item 53	-1.73	3.99
	Item 54	-1.69	3.54
8. Kinetic object segmentation	Item 25	-2.76	2.37
	Item 26	-2.51	2.78
	Item 27	-2.58	3.56

Subtest	Item	Difficulty	Discrimination
	Item 28	-2.76	3.04
	Item 29	-2.44	1.80
9. Biological motion	Item 40	-1.02	1.25
	Item 41	-1.37	2.61
	Item 42	-1.19	2.95
	Item 43	-2.02	2.69
	Item 44	-2.55	1.58
10. Dot counting	Item 55	-2.28	1.77
	Item 56	-2.67	1.50
	Item 57	-1.93	2.30
	Item 58	-1.62	1.29
	Item 59	-2.21	1.88
11. Figure-ground segmentation	Item 20	-2.88	2.29
	Item 21	-2.46	1.60
	Item 22	-2.20	2.48
	Item 23	-2.14	2.32
	Item 24	-2.72	2.82
12. Embedded figure detection	Item 65	-3.87	1.37
	Item 66	-1.46	1.38
	Item 67	-1.58	1.18
	Item 68	-1.54	1.99
	Item 69	-1.44	1.45
13. Recognition of missing part	Item 45	-2.21	1.43
	Item 46	-2.10	1.74
	Item 47	-2.49	2.36
	Item 48	-2.16	1.91
	Item 49	-2.70	2.10
14–15. Object recognition in a scene	Item 75	-2.93	1.61
	Item 76	-2.60	2.51
	Item 77	-2.40	2.19
	Item 78	-2.97	2.12
	Item 79	-2.52	2.04

Table S2

Standardised Factor Loadings of Final Model With Subtests as Factors and Subtests 4, 5, and 6 Combined Into One Factor.

Item	1. Fine shape discrimination	2. Shape ratio discrimination (Efron)	3. Dot lattices	4-6. RFP	7. Global motion detection	8. Kinetic object segmentation	9. Biological motion	10. Dot counting	11. Figure-ground segmentation	12. Embedded figure detection	13. Recognition of missing part	14–15. Object recognition in a scene
Item 3	0.66											
Item 2	0.62											
Item 1	0.62											
Item 4	0.61											
Item 0	0.60											
Item 19		0.58										
Item 17		0.53										
Item 15		0.52										
Item 16		0.48										
Item 18		0.47										
Item 7			0.73									
Item 9			0.69									
Item 8			0.66									
Item 5			0.66									
Item 6			0.55									
Item 37				0.84								
Item 32				0.76								
Item 35				0.75								
Item 30				0.73								
Item 38				0.73								
Item 12				0.72								
Item 39				0.69								
Item 14				0.68								
Item 11				0.67								
Item 13				0.63								
Item 33				0.63								
Item 36				0.62								
Item 34				0.61								
Item 10				0.60								
Item 31				0.57								
Item 53					0.93							
Item 51					0.88							
Item 54					0.87							
Item 52					0.85							
Item 50					0.81							
Item 28						0.96						
Item 27						0.89						
Item 25						0.79						
Item 26						0.75						
Item 29						0.70						
Item 42							0.86					
Item 44							0.84					
Item 43							0.76					
Item 41							0.72					
Item 40							0.59					
Item 59								0.78				
Item 56								0.74				
Item 55								0.69				
Item 57								0.66				

Table S3

Latent Factor Correlation Matrix of Final Model With Subtests as Factors and Subtests 4, 5, and 6 Combined Into One Factor.

Subtest	1. Fine shape discrimination	2. Shape ratio discrimination (Efron)	3. Dot lattices	4-6. RFP	7. Global motion detection	8. Kinetic object segmentation	9. Biological motion	10. Dot counting	11. Figure-ground segmentation	12. Embedded figure detection	13. Recognition of missing part
1. Fine shape discrimination											
2. Shape ratio discrimination (Efron)	0.81										
3. Dot lattices	0.51	0.40									
4-6. RFP	0.80	0.75	0.55								
7. Global motion detection	0.40	0.27	0.29	0.47							
8. Kinetic object segmentation	0.71	0.68	0.43	0.77	0.43						
9. Biological motion	0.48	0.37	0.28	0.50	0.33	0.48					
10. Dot counting	0.66	0.63	0.41	0.60	0.34	0.63	0.38				
11. Figure-ground segmentation	0.66	0.60	0.49	0.75	0.35	0.62	0.43	0.57			
12. Embedded figure detection	0.69	0.54	0.41	0.63	0.42	0.49	0.42	0.48	0.58		
13. Recognition of missing part	0.67	0.63	0.42	0.63	0.37	0.55	0.35	0.53	0.65	0.56	
14–15. Object recognition in a scene	0.48	0.34	0.34	0.54	0.24	0.53	0.31	0.37	0.56	0.40	0.49