

ORIGINAL ARTICLE

OPEN

# Development of the Contact Lens User Experience: CLUE Scales

R. J. Wirth\*, Michael C. Edwards\*, Michael Henderson†, Terri Henderson†, Giovanna Olivares‡, and Carrie R. Houts\*

## ABSTRACT

**Purpose.** The field of optometry has become increasingly interested in patient-reported outcomes, reflecting a common trend occurring across the spectrum of healthcare. This article reviews the development of the Contact Lens User Experience: CLUE system designed to assess patient evaluations of contact lenses. CLUE was built using modern psychometric methods such as factor analysis and item response theory.

**Methods.** The qualitative process through which relevant domains were identified is outlined as well as the process of creating initial item banks. Psychometric analyses were conducted on the initial item banks and refinements were made to the domains and items. Following this data-driven refinement phase, a second round of data was collected to further refine the items and obtain final item response theory item parameters estimates.

**Results.** Extensive qualitative work identified three key areas patients consider important when describing their experience with contact lenses. Based on item content and psychometric dimensionality assessments, the developing CLUE instruments were ultimately focused around four domains: comfort, vision, handling, and packaging. Item response theory parameters were estimated for the CLUE item banks (377 items), and the resulting scales were found to provide precise and reliable assignment of scores detailing users' subjective experiences with contact lenses.

**Conclusions.** The CLUE family of instruments, as it currently exists, exhibits excellent psychometric properties. (Optom Vis Sci 2016;93:801–808)

Key Words: psychometrics, patient-reported outcomes, item response theory, scale development, contact lenses

Evaluating the suitability of soft, disposable contact lenses (CLs) involves measurement outside of just visual acuity tests and slit-lamp examinations. Despite the outstanding CL products that are available in the market today, global estimates

indicate that, of patients who discontinue CL use, over 40% report dropping out due to discomfort.<sup>1</sup> A more recent retrospective chart audit of patients new to CL wear found poor vision as the main reason for discontinuing CL use.<sup>2</sup> Thus, patients' subjective impressions while wearing CLs are important outcomes to assess in developing new CL products and, in the clinical setting, retaining patients in CLs. These experiences may include how comfortable the lenses are, how satisfied the patient is with the vision provided, how the lenses handle, the convenience of the required cleaning routine, or even cost.

Often researchers evaluate subjective patient outcomes using long questionnaires followed by a series of single-item analyses. This can be problematic for at least three reasons. First, it has been shown that individual items are typically predominately comprised of error (i.e. noise) and thus any single item does a poor job reflecting an individual's "true" experience.<sup>3</sup> Second, analyzing an often large number of individual items results in a large decrement in statistical power (if a more stringent alpha is used to statistically control for multiple testing) or a large increase in Type I error (if typical alphas are used without consideration for the multiple testing). Third, it is

\*PhD

†MS

‡OD, FAAO

Vector Psychometric Group, LLC (RJW, MCE, CRH), Chapel Hill, North Carolina; SAS Institute, Inc. (MH), Cary, North Carolina; and Johnson & Johnson Vision Care, Inc. (TH, GO), Jacksonville, Florida.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site ([www.optvissci.com](http://www.optvissci.com)).

This is an open-access article distributed under the terms of the Creative Commons Attribution-Non Commercial-No Derivatives License 4.0 (CCBY-NC-ND), where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially.

The Contact Lens User Experience™ (CLUE™) measurement system which includes the item statements or questions, the item response options, and item response theory based parameter values, published as SDC to this article, is copyrighted by Johnson & Johnson Vision Care, Inc.

often difficult, if not impossible, to capture the complexity and breadth of the intended construct with a single item.

Given these issues, the development of measures that can produce reliable scores and maintain, if not increase, statistical power without inflating false discovery rates is highly desirable. The availability of an assessment capable of producing reliable scores and supporting valid inferences related to patient impressions of important CL characteristics could be a great benefit to researchers who are interested in measuring these subjective outcomes. In addition, the availability of psychometrically rigorous scales can provide greater confidence when making critical decision such as bringing new products or treatments to market. Clinically, such a tool could aid practitioners in providing each patient with the best possible CLs for their needs and assuring the continuing satisfaction of their patients throughout their lifetime of CL wear.

Although there are existing instruments in optometry developed using modern scale development techniques, we were unable to find any instrument or collection of instruments to meet our specific needs. For instance, the Contact Lens Dry Eye Questionnaire<sup>4</sup> is intended for use in assessing eye dryness associated with CL use, neglecting most other possible aspects of the CL wearer's subjective experience. The Quality of Vision Questionnaire<sup>5</sup> covers a range of visual symptoms patients may experience but is not specific to contact lenses and does not address other aspects of the CL experience beyond just a lack of symptoms, such as poor vision due to CLs moving. The Ocular Comfort Index<sup>6</sup> assesses eye comfort but is, again, not specific to CL evaluation and, as such, fails to include any CL-specific aspects that may affect comfort, such as lenses feeling bulky or stiff or a persistent awareness of the lenses. The Contact Lens Impact on Quality of Life Questionnaire (CLIQ)<sup>7</sup> assesses a wide range of topics related to contact lens use, but provides only a single overall score, which does not allow for differentiation among the subareas that inform CL-related QOL, such as comfort or vision, which was desired.

The commercial success of a CL product is dependent on how well the new lens material and/or design is able to address patients' needs. The current state of CL products is such that improvements are often incremental in nature. Consider a small design tweak to a CL model—a single overall score, such as provided by CLIQ, may not be sufficiently sensitive to detect a small (but noteworthy) improvement in comfort that this modification affects. That is, the incremental improvement in comfort would likely be masked by the various other static aspects (vision, pricing, convenience, self-confidence, etc.) of the CLs also incorporated into that single score. Given the lack of suitable scales to assist in the development of new CL products, the current project was designed with the goal of capturing the primary domains responsible for forming patients' opinions specific to CL products, as opposed to CL wear in general which may include other factors such as convenience or self-confidence, and then using that information to develop an initial item bank for each identified domain.

This paper outlines the development of the Contact Lens User Experience: CLUE measures. The development process employed comprehensive qualitative methods in the early stages of concept elucidation and definition and used rigorous statistical principles and methods to analyze the resulting data. This process is similar to the patient-reported outcome development steps recommended

by the FDA<sup>8</sup> and standard scale development practices in the field of modern psychometrics.<sup>9</sup> The use of modern psychometric techniques for this project is important. A primary goal of this project was always to construct large item banks, rather than a fixed-length scale for each identified domain. Such item banks, when established using modern psychometrics, allow for assessments to be tailored to individuals (such as in computerized adaptive testing), to mode of administration (e.g. paper and pencil versus electronic data collection), or to a predetermined length (such as maximizing the reliability of a fixed-length 15-item assessment), all while providing domain scores that are directly comparable to each other regardless of the specific items an individual or study group sees. That is, we will present evidence for the psychometric properties of several domain-specific items banks. These item banks can be large (one containing close to 200 items), but it was never intended that subjects would see all 200 items; rather, CLUE was developed for continued use with modern psychometrics, allowing researchers or practitioners to create assessments suited to their current informational goals. The CLUE assessments were developed using a sample of healthy adult, soft, disposable CL wearers and are specifically designed to assess an individual's subjective experiences with CLs on domains empirically identified as critical to the evaluation of any currently worn CL product.

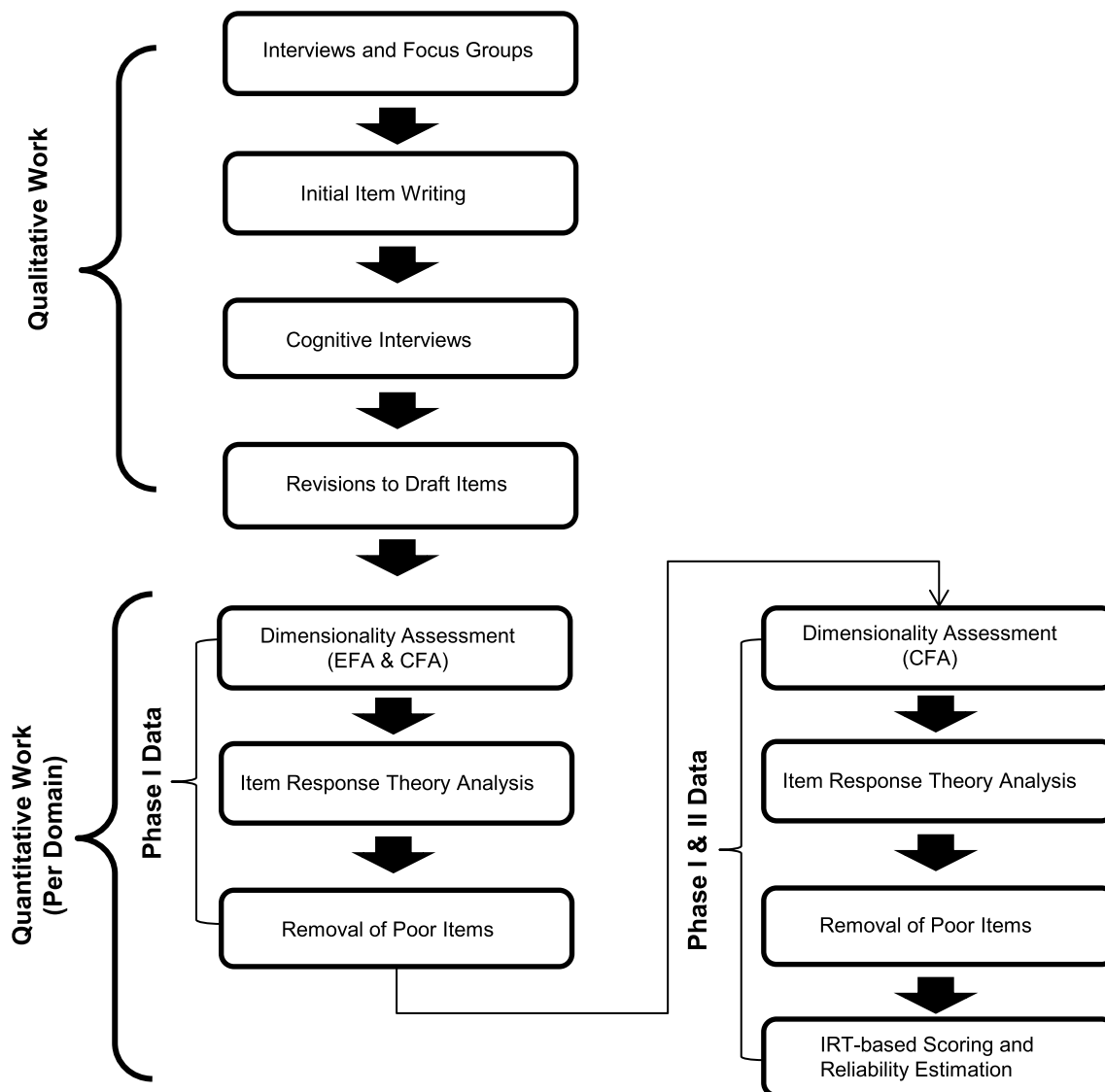
## METHODS

Qualitative research involved the use of one-on-one interviews, focus groups, and cognitive interviews using a limited number of patients. Quantitative research involved obtaining data from subjects in two separate rounds of data collection, across several different sites, in which patients were asked to respond to candidate items from the developing measures. For clarity, an overview of both the qualitative and quantitative research methods employed is summarized in Fig. 1.

## Participants

In this project, the development process focused on all healthy adult, soft, disposable CL wearers receiving at least minimal care from an optometrist (enough to have a current prescription) and who wear CLs in their daily, real-world environments. Participants in the qualitative research (N = 86 across all interviews, focus groups, and cognitive interviews) were from both the US and the UK. The qualitative research was conducted in the US and UK with English-speaking subjects wearing their habitual soft contact lenses, and all participants were considered to be in good health (no visual or health problems that would impact their responses). Patients in the qualitative research included both males and females and ranged in age from 17 to 66. Patients used a variety of soft contact lenses (spherical, toric, multifocal, daily disposable, monthly, etc.). After initial item generation, 25 subjects were recruited for structured cognitive interviews of the developing item banks.

Participants in the quantitative research studies were also recruited from both the US and the UK. Participants were eligible for inclusion if they were at least 18 years of age and no more than 55 years of age, had visual acuity best correctable to 20/30 or better for each eye, read and signed the statement of informed consent,



**FIGURE 1.**

Overview of qualitative and quantitative work conducted in the development of the CLUE assessment system.

appeared able and willing to adhere to the instructions of the study, and were a current soft contact lens wearer for at least 1 week. Subjects were excluded from the study if they had any conditions/took medications that would interfere with CL wear, any damage/abnormalities of the cornea, clinically significant (grade 3 or 4) tarsal abnormalities or bulbar hyperemia, any ocular infection, were pregnant or lactating, had diabetes, had any infectious diseases and/or contagious immunosuppressive diseases, or had a history of chronic eye disease. The clinical protocols were approved by Sterling Institutional Review Board (Atlanta, GA) and performed in compliance with the Declaration of Helsinki. Informed consent was obtained from each study participant before enrollment.

Subjects in phase I of the quantitative portion of the study were required to come for a single visit 1 week after recruitment to complete the questionnaire and undergo biomicroscopic evaluation while wearing their habitual CLs. For phase II, subjects were required to come for a total of two visits to complete the questionnaire while wearing their habitual CLs: a baseline visit and a follow-up visit 4 to 10 days later. Patients were not

allowed to enroll in both phases of the quantitative studies. Soft contact lenses were habitually worn by most of the cohort; however, specialty brands were worn as well. Table 1 provides a summary of several demographic and descriptive variables, both by data collection round and combined, as the final item response theory (IRT) analyses were performed on this combined sample.

### Qualitative Research

The initial domain identification and content collection process involved conducting structured one-on-one interviews and semi-structured focus groups. The information gleaned from the interviews guided the focus group discussions. Each focus group was conducted by a single qualitative researcher after informed consent was obtained.

After the interviews and focus groups, an initial round of item writing was undertaken. The developed items were presented to a new group of subjects for a preliminary in-depth assessment of the item characteristics. The cognitive interviews were completed

**TABLE 1.**

Demographic and descriptive variable summary of the quantitative samples

Variable	N	Mean	SD	Min	Median	Max
<b>Phase I</b>						
Sex	461					
Female	330					
Male	129					
Missing	2					
Age	461	33.30	10.65	18	33	65
Cylinder	920	-0.56	0.73	-7.5	-0.25	0
Sphere	920	-3.33	2.57	-16.5	-3	6
<b>Phase II</b>						
Sex	525					
Female	389					
Male	136					
Age	525	33.67	10.76	18	33	56
Cylinder	994	-0.56	0.73	-8	-0.5	0
Sphere	994	-3.02	2.69	-14.5	-3	7
<b>Combined phase I and II</b>						
Sex	986					
Female	719					
Male	265					
Missing	2					
Age	986	33.50	10.71	18	33	65
Cylinder	1914	-0.56	0.73	-8	-0.25	0
Sphere	1914	-3.17	2.64	-16.5	-3	7

one-on-one by trained qualitative researchers after informed consent was obtained. Using a structured interview, each subject was asked their opinion about the appropriateness of the presented items, whether any item required additional qualifications, the language used in the items, the response formats, complexity, and the clarity of the items.

### Quantitative Analyses<sup>a</sup>

The items available after the qualitative research were presented to the phase I and phase II quantitative study patients. Data manipulation and descriptive statistics of the item responses were completed using SAS 9.0. Initial analyses included the examination of item responses for ceiling or floor effects, constant skips, or other potentially problematic response patterns.

### Exploratory Factor Analysis (EFA)

EFAs were performed within domain, as the primary interest was in establishing that each of the three domain-specific item banks (Comfort, Vision, and Handling) was primarily assessing one construct. All EFAs were estimated using unweighted least squares and polychoric correlations in Mplus 4.2.<sup>10</sup>

### Confirmatory Factor Analysis (CFA)

All CFAs were conducted using diagonally weighted least squares estimation and polychoric correlations in Lisrel 8.8.<sup>11</sup> In the CFA

<sup>a</sup>An overview of the quantitative methods employed here is beyond the scope of this paper, but the interested reader is referred to Wirth and Edwards<sup>19</sup> for an accessible didactic article on modern statistical scale development techniques.

framework, there are statistics available which indicate the extent to which a particular model accounts for correlations among the observed item responses. The three domains were designed to each represent one construct, so the target model in each CFA was a one-factor model. To assess the model fit, the root mean square error of approximation (RMSEA) and the comparative fit index<sup>12</sup> (CFI) were used. RMSEA values (lower values indicate better model fit) are considered acceptable below 0.1, but published standards make further distinctions below 0.08 and 0.05.<sup>13</sup> The CFI (ranges from 0 to 1, with higher values indicating better model fit) is generally viewed as supporting model fit if it is greater than 0.95.<sup>14</sup>

### IRT Analyses

All IRT analyses were conducted using the graded response model (GRM)<sup>15</sup> in Multilog 7.0.3.<sup>16</sup> The GRM was chosen due to the polytomous, ordered response categories of the items (primarily five-category level-of-agreement response options) and the belief that some items would be better indicators (that is, have higher slope parameter values) of the measured constructs than other. Because of our a priori belief that the slopes would differ across items (i.e. some items would be more strongly related to the construct than others), a model constraining all slopes to be equal, as is done within Rasch-consistent models, was deemed inappropriate for this project. The unidimensional GRM results in a single slope parameter estimate for each item, which provides a numerical value of how well the item measures the construct, and, for an item with  $m$  response categories,  $m - 1$  severity parameters, which link the response options to the underlying latent construct.

## RESULTS

### Qualitative Research

Qualitative analyses of the initial interviews and focus groups identified 15 topics which appeared important in measuring characteristics of CLs that inform a wearer's subjective evaluation of a CL product. Upon detailed evaluation, these 15 areas could be seen to fall into 3 overarching content domains: comfort, handling, and vision. Table 2 presents the 15 identified topics and the overarching domains to which each topic was assigned.

To reduce possible respondent burden, items were written to conform to a standardized format. The statements were kept short and common words were used when possible to maintain readability. Using the common Flesch-Kincaid<sup>17,18</sup> statistics, the items were found to have an average readability score of 68.7 and a 6.7 grade level. To minimize inter-individual variability in recall time frames, draft items were edited when possible to refer to specific recall periods (e.g. "5 minutes") as opposed to general recall periods (e.g. "a few minutes"). The response options chosen for the majority of CLUE items consisted of a 5-point Likert-type response format that included the anchors "strongly disagree," "disagree," "neither agree nor disagree," "agree," and "strongly agree." Due to item content, Yes/No responses were appropriate for three of the items (e.g. "Every day, it took more than one attempt to insert the lenses").

A total of 291 items (89 Comfort, 49 Handling, and 153 Vision items) were presented to patients for review. Cognitive interviews

**TABLE 2.**

Topics identified as important in CL wear by patients and practitioners during qualitative work, presented by researcher-constructed domains

Comfort	Handling	Vision
General comfort	General handling	General vision
Lens awareness	Quality/ease of packaging	Visual quality
Level of comfort throughout the day	Preparing the lenses for insertion	Visual clarity
Symptoms of discomfort (caused by lenses)	Insertion of the lenses	Sight during activities of daily living (including time of day)
“feel” of the lenses during daily living activities	Removal of the lenses	
	Cleaning the lenses	

found that individuals could follow the item structure, response format, and instructions. Based on the composite results of the cognitive interviews, minor edits were made (when possible) to items noted as being worded negatively, difficult to understand, or with a time frame reference that was deemed too vague. Although no major changes were made to the items or the item content, 39 items were dropped; these dropped items were primarily slight wording variations of items examined for subject wording preference.

### Quantitative Analyses

The 252 items remaining after the cognitive interviews were supplemented with an additional 32 items used to study the impact of different response options (results not reported here), for a total of 284 items presented to patients in phase I of the quantitative data collection. As described in the Methods section, these items were submitted to EFA, CFA, and IRT analysis. As the phase I analyses were provisional in the scale-building process, results are only generally summarized (detailed methods and results are available from the first author upon request). Results from the EFAs suggested that each domain (Comfort, Handling, and Vision) was composed of one primary construct (that is, they were essentially unidimensional scales). The results of the CFAs also supported a single domain underlying each item set, with adequate fit values being obtained for a unidimensional model within each domain. Given the results of the EFAs and CFAs, it was determined that there was sufficient evidence to support the use of the unidimensional GRM IRT model within each of the three domains.

Out of the 284 items which were analyzed in the phase I IRT calibrations, only 11 items (3 in Handling, 8 in Vision) had slope estimates which were above 4.0, a value typically considered the maximum for reasonable estimates.<sup>19</sup> These items, while candidates for removal, were retained for the planned subsequent study to examine if additional data would stabilize the parameter estimates.

There were also items which seemed to have very little relation to their intended domain (that is, exhibited low estimated slope values). These poorly performing items (39 total, 20 in Comfort, and 19 in Vision) were removed from the banks and were not included in any further data collection or analyses.

At the end of phase I, a total of 245 tested CLUE items remained (84 Comfort, 44 Handling, and 117 Vision). Using an IRT method for visually examining the quality of measurement over the range of the construct, it was apparent that each domain could benefit by the addition of more difficult-to-endorse items (e.g. items for very comfortable CL experiences). Thus, additional items (138) were generated before phase II in an attempt to fill this need.

In phase II, item responses were collected on a total of 383 items, split between 2 forms containing 276 and 275 items, respectively, to reduce respondent burden. To further reduce respondent burden, each form was further divided into two subsections—patients completed only one subsection of their assigned form per visit. Before the beginning of the phase II primary psychometric analyses, all items were reviewed for face validity and domain specificity and correlations among items within each domain were examined. During this review, it became apparent that the Handling domain was more accurately represented as two distinct factors: Handling and Packaging. Twelve of the original Handling items were reclassified as Packaging items. Although packaging is an important part of the users' experience with CLs, it was not the intended focus of the Handling item development. This split enables researchers to assess patient satisfaction with packaging when appropriate and receive a score related to just this aspect of the product. Splitting Handling into two domains resulted in four domains assessed by CLUE items during the primary phase II psychometric analyses: Comfort, Handling, Packaging, and Vision.

Primary psychometric analyses began with an examination of item level response frequencies which found that not all response options were used for 142 of the items. For these items, categories with fewer than five people were collapsed into the adjacent response category before factor and IRT analyses. The dimensionality of the Comfort, Handling, Packaging, and Vision item banks were examined using CFA. A one-factor model was examined for each domain by form/subsection. This resulted in a total of 16 models (1 model for Form A1, A2, B1, and B2 for each domain). The CFA results (see Table 3) supported a single domain underlying each item set. Only two of the models had RMSEA values above the 0.10 cutoff. Considered as a group, the CFA results provided strong evidence that each of the four domains (Comfort, Handling, Packaging, and Vision) could be adequately described with one latent factor. The CFA results provided sufficient evidence to move forward with the planned IRT analyses, conducted using the combined phase I and phase II samples, as noted earlier.

The Comfort domain had a total of 128 items in the final IRT calibration. All of these items had reasonable parameter estimates and appeared to be providing useful information about the perceived comfort of the CLs. More specifically, Comfort items were found to have slope estimates between 0.66 and 3.30 ( $M_a = 1.69$ ,  $SD_a = 0.49$ ). The severity parameter estimates suggested that the Comfort items covered a broad range of comfort experiences ( $M_{b1} = -2.94$ ,  $SD_{b1} = 0.78$ ;  $M_{b2} = -1.36$ ,  $SD_{b2} = 0.60$ ;  $M_{b3} = -0.30$ ,  $SD_{b3} = 0.80$ ;  $M_{b4} = 1.32$ ,  $SD_{b4} = 0.64$ ) ranging from  $-6.11$  to  $3.70$ . Online Appendix Table A1 (available at

**TABLE 3.**

CFA model fit (RMSEA and CFI) values for the Comfort, Handling, Packaging, and Vision domains by form

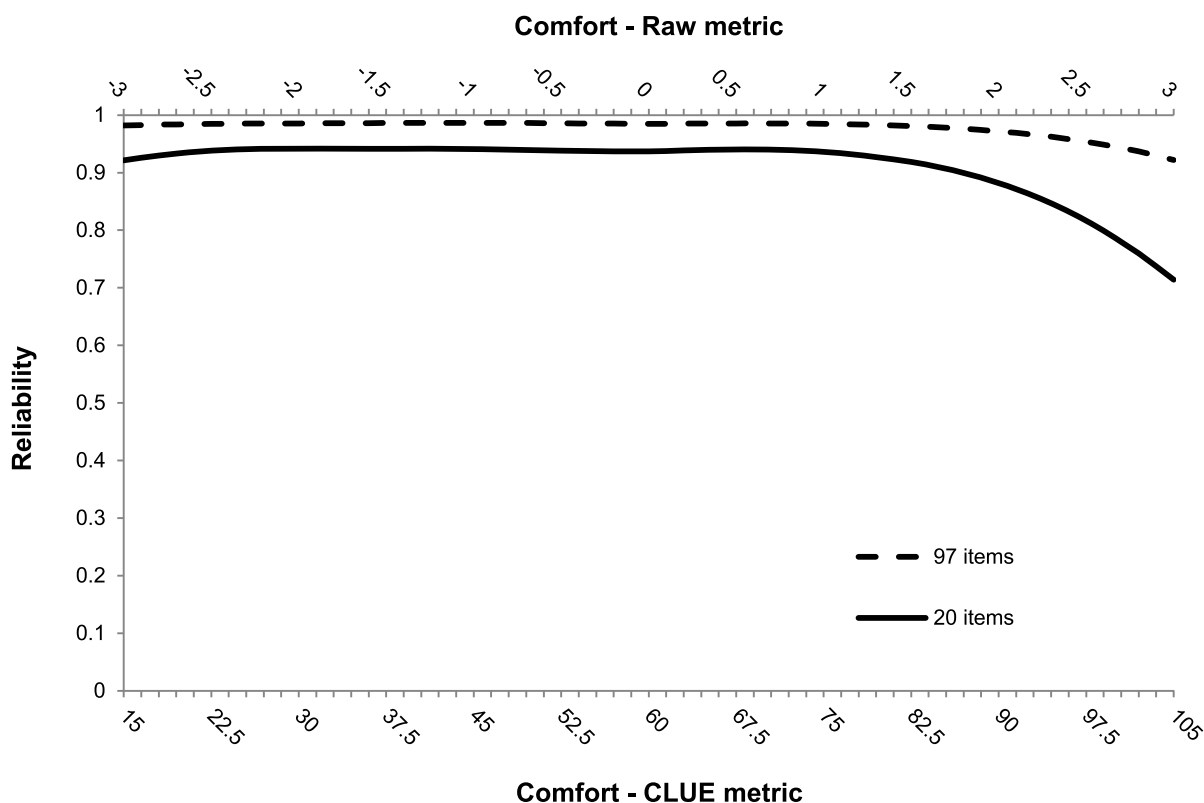
Domain	Form	RMSEA	CFI
Comfort	A1	0.08	0.97
	A2	0.11	0.96
	B1	0.09	0.97
	B2	0.09	0.95
Handling	A1	0.05	0.99
	A2	0.05	0.99
	B1	0.05	0.99
	B2	0.08	0.97
Packaging	A1	0.09	0.96
	A2	0.00	1.00
	B1	—	—
	B2	0.04	0.99
Vision	A1	0.08	0.98
	A2	0.10	0.97
	B1	0.11	0.96
	B2	0.02	1.00

The model for Packaging, Form B1 was saturated, making fit assessment impossible.

<http://links.lww.com/OPX/A244> provides 20 example items from the CLUE Comfort item bank, with both item text and estimated factor loadings. These items were selected to demonstrate the range of factor loadings, including both the item with the highest and lowest factor loading in the bank and a sampling of item content. Factor loadings are reported to provide a metric likely

more familiar to readers, with values that may range from 0 to 1 and higher values indicating items more closely related to the construct, overall Comfort in this case.

The scale scores are, in raw form, from a standard normal distribution (mean = 0, SD = 1) but have been developed such that they can be calculated by taking the IRT-based raw modal a posteriori (MAP) score, multiplying it by 15 and then adding 60. This gives each domain a score that roughly ranges from 0 to 120 with a mean of 60 and a standard deviation of 15. The mean of these scores reflects the average level of the construct (e.g. Comfort) in the population of healthy adult, soft, disposable CL wearers, with higher scores indicating more experienced satisfaction with the measured construct. A test constructed using a subset of 97 items found that Comfort scores had high reliability (>0.92) over a broad range of experienced comfort (see Fig. 2). For much of the scale, between -3 and 1.5 in the raw score metric, score reliabilities are above 0.98, and at the higher end, estimated reliability drops somewhat. This suggests that the upper end of the perceived Comfort domain (i.e. those who are very comfortable) is not as reliably measured as the lower end (i.e. those who are very uncomfortable). This is a trend that is repeated in all the constructs measured during this study. Also included in Fig. 2 is a reliability plot in for an assessment made of 20 randomly chosen CLUE Comfort items. As can be seen, even for the 20-item version, the reliability is still quite good, over 0.90 for much of the range; however, the noted drop at the high end of the continuum is much more pronounced for this shortened form, although even at its lowest point the expected reliability is still above 0.70, a commonly used minimum value.

**FIGURE 2.**

Reliability plot for 2 subsets of CLUE Comfort items.

The Handling domain began phase II with a total of 48 items. However, five of these items were deleted due to very low slope estimates. For example, the item “I have ripped only 1 of these lenses” had a slope-parameter estimate of 0.09. The other four Handling items removed had slopes less than 0.27. This resulted in a final Handling item bank of 43 items. The top portion of online Appendix Table A2 (available at <http://links.lww.com/OPX/A244>) provides 10 example items from the CLUE Handling item bank, with both item text and estimated factor loadings. These items were selected to demonstrate the range of factor loadings, including both the item with the highest and lowest factor loading in the bank, and a sampling of item content. Handling items were found to have lower slope estimates, on average, compared to Comfort items ( $M_a = 1.97$ ,  $SD_a = 0.65$ ) and ranged from 0.53 to 3.04. The severity parameter estimates ranged from -6.41 to 2.96 and were found to cover a wide range of handling experiences on average ( $M_{b1} = -2.64$ ,  $SD_{b1} = 1.09$ ;  $M_{b2} = -1.54$ ,  $SD_{b2} = 0.71$ ;  $M_{b3} = -0.25$ ,  $SD_{b3} = 1.00$ ;  $M_{b4} = 1.28$ ,  $SD_{b4} = 0.62$ ). Using a subset of 31 Handling items, the resulting scale produces very reliable scores. From 3 standard deviations below the mean to 1.4 above the average score, reliabilities are 0.95 or higher. There is a steep drop-off in reliability as scores move above 1.4. Still, scores between 1.4 and 1.9 have reliabilities over 0.9 and scores between 1.9 and 2.4 have reliabilities over 0.8.

Because Packaging was not an original focus of this scale development, the domain consisted of only 12 items. However, these 12 items were found to have slope parameter estimates ranging from 0.54 to 2.61 ( $M_a = 1.62$ ,  $SD_a = 0.62$ ), suggesting they all provided useful information regarding users' package handling experiences. The severity parameter estimates ranged from -5.00 to 1.39 ( $M_{b1} = -2.98$ ,  $SD_{b1} = 0.81$ ;  $M_{b2} = -1.69$ ,  $SD_{b2} = 0.57$ ;  $M_{b3} = -0.21$ ,  $SD_{b3} = 1.35$ ;  $M_{b4} = 1.07$ ,  $SD_{b4} = 0.22$ ) suggesting that although the items may be informative, they do not provide much information across a broad range of packaging experiences. The lower portion of online Appendix Table A2 (available at <http://links.lww.com/OPX/A244>) provides both item text and estimated factor loadings for 10 example items from the CLUE Packaging item bank. The Packaging items produce scores with reliabilities over 0.8 from 3 standard deviations below the mean to 1.9 above. Over much of this range, the score reliabilities are closer to 0.9 than to 0.8. Again, there is a considerable drop-off in score reliability as satisfaction with packaging increases beyond 1.9.

The Vision domain was the largest of the 4, with 195 items. One of these items (*There were some activities that I preferred to do after inserting these lenses*) was deleted due to a low slope estimate ( $a = 0.40$ ). The resulting Vision item bank consisted of 194 items. Online Appendix Table A3 (available at <http://links.lww.com/OPX/A244>) provides 20 example items from the CLUE Vision item bank, reporting both item text and estimated factor loadings. These items were selected to demonstrate the range of factor loadings, including both the item with the highest and lowest factor loading in the bank, and a sampling of item content. The Vision items were found to have slope estimates between 0.66 and 3.40 ( $M_a = 2.14$ ,  $SD_a = 0.42$ ). Much like the Comfort item set, the severity parameter estimates for the Vision item set suggested that the Vision items covered a broad range of perceived vision ranging from -5.79 to 3.42 ( $M_{b1} = -2.44$ ,  $SD_{b1} = 0.55$ ;  $M_{b2} = -1.55$ ,

$SD_{b2} = 0.45$ ;  $M_{b3} = -0.36$ ,  $SD_{b3} = 0.95$ ;  $M_{b4} = 0.86$ ,  $SD_{b4} = 0.36$ ). Constructing a test using a subset of 122 Vision items yielded highly reliable scores. The reliability of the scores between 3 standard deviations below the mean and 1.7 above all exceed 0.98. The least reliable score, 3 standard deviations above the mean, still maintained a reliability of 0.86.

## DISCUSSION

The initial qualitative work, which is the foundation for developing validity arguments, provided great insight and enabled us to understand the measured constructs as the subjects do and assess them accordingly. CFAs within each of the domains supported the contention that a single construct was being measured by each domain. IRT item parameters were successfully estimated for all candidate items. Some items, based on poor performance as indicated by low slope value estimates, were eliminated from the final CLUE item bank.

Overall, the final CLUE assessment comprised 4 domains (Comfort, Vision, Handling, and Packaging) and a total of 377 items (128, 194, 43, and 12 items in each domain, respectively). The expected reliability of the scores obtained from all four domains was found to be above minimal acceptable levels, with expected reliabilities for some scales reaching over 0.90 for a large range of the measured continuum. However, the CLUE scores, across all four domains, do tend to exhibit a noticeable drop-off in reliability at the extreme positive end of the continuum.

Using methods consistent with standards in scale development, the creation of the CLUE item banks provides researchers and practitioners with a flexible platform which, to date, appears able to produce very reliable and valid scores in the domains of Comfort, Vision, Handling, and Packaging. Although the item banks contain a great many items, they were developed with the intention of either administering the scales adaptively using computerized adaptive testing or creating forms that provide an appropriate level of reliability for the intended use/purpose; either method would greatly reduce the number of items a patient would be required to respond to while still providing scores that are directly comparable, regardless of the exact items a given patient sees. This degree of flexibility, while maintaining comparability, highlights one of the many benefits of using IRT as a psychometric model for scale development.

Understanding the patient's subjective evaluation in detail is often a critical aspect of improving the quality of products (like CLs). It is obvious the importance and value that CLUE can bring to new product development for a manufacturer. New CL products require that the patient needs and insights be identified and then incorporated into a lens material and/or design. Therefore, a sensitive instrument that measures the subjective patient experience, such as CLUE, can be invaluable in the development of new CL products.

Clinically, to maintain a healthy and growing CL practice, understanding the reasons for CL discontinuation and patient satisfaction would allow a practitioner to identify deficient aspects of the current lenses, make informed adjustments to a lens better suited to the patient, and thereby maintain the patient in CLs. Although developed primarily for corporate product assessment, we believe that CLUE could benefit practitioners as well. CLUE

could be administered easily to all CL patients, either before their office visit or while in the waiting room before the examination. To this end, CLUE is available as a paper-and-pencil assessment using items selected by the practitioner or via statistical criteria. Additionally, CLUE can be administered either on a practitioner-provided terminal/device or on a patient's own device, such as a smart phone or tablet. A computer-based administration system allows for CLUE to be administered either as a fixed set of items or adaptively using state-of-the-art computerized adaptive testing software. The practitioner would use the CLUE results to evaluate the patient's subjective CL experience to normative data and compare the patient's current satisfaction against the CL population and/or their previous visit's scores. The objective of using CLUE in this manner would be to ascertain whether the patient's CL experience is at or above the average scores, particularly in the domains that appear to matter most, Comfort and Vision. If the scores are below average or average, a change in design, material, or brand may be warranted to provide the patient with the best CL experience. Those interested in CLUE for use in their practice/research may contact [Info@VPGcentral.com](mailto:Info@VPGcentral.com).

As with any study or development effort, there are limitations. Although the recommended qualitative steps were used to begin the process, and there is a substantial amount of construct-related validity evidence, more is recommended. Additional studies have been and are continuing to be conducted which will allow a fully articulated argument of appropriate uses and inferences for scores from the four constructs being measured. To some extent, the gathering of validity evidence is a never-ending task, but early results have been promising as the work with CLUE continues.

Lastly, it is important to note that the work described here may or may not prove reflective of other populations. This is not a limitation of the work reported, but rather a fact of all scale development. As researchers seek to extend the use of CLUE beyond the healthy population of adults who routinely wear CLs it was developed in, much of the process will need to be repeated anew to ensure that, at the end of the process, the intended constructs are indeed the ones being measured.

## ACKNOWLEDGMENTS

*Vector Psychometric Group, LLC received financial compensation from JJVCI, Inc. for the study design and analyses presented in this manuscript.*

*The authors thank Youssef Toubouti for his thoughtful comments on the manuscript.*

*Received September 1, 2015; accepted May 3, 2016.*

## APPENDIX

Appendix Tables A1 to A3, which provide item text and factor loadings for a large subset of the available CLUE items, in tables separated by domain, are available at <http://links.lww.com/OPX/A244>.

## REFERENCES

- Rumpakis J. New Data on Contact Lens Dropouts: An International Perspective. Available at: [http://www.reviewofoptometry.com/content/d/contact\\_lenses\\_\\_\\_and\\_\\_\\_solutions/c/18929/](http://www.reviewofoptometry.com/content/d/contact_lenses___and___solutions/c/18929/). Accessed November 10, 2015.
- Sulley A, Young G, Hunt C. Factors in the success of new contact lens wearer retention. *Optom Vis Sci* 2014;91:E-abstract 145020.
- Churchill GA, Jr. A paradigm for developing better marketing constructs. *J Mark Res* 1979;16:64–73.
- Nichols JJ, Mitchell GL, Nichols KK, Chalmers R, Begley C. The performance of the contact lens dry eye questionnaire as a screening survey for contact lens-related dry eye. *Cornea* 2002;21:469–75.
- McAlinden C, Pesudovs K, Moore JE. The development of an instrument to measure quality of vision: the Quality of Vision (QoV) questionnaire. *Invest Ophthalmol Vis Sci* 2010;51:5537–45.
- Johnson ME, Murphy PJ. Measurement of ocular surface irritation on a linear interval scale with the ocular comfort index. *Invest Ophthalmol Vis Sci* 2007;48:4451–8.
- Pesudovs K, Garamendi E, Elliott DB. The Contact Lens Impact on Quality of Life (CLIQ) Questionnaire: development and validation. *Invest Ophthalmol Vis Sci* 2006;47:2789–96.
- U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER), Center for Devices and Radiological Health (CDRH). Guidance for Industry Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. Washington, DC: Office of Communications, Division of Drug Information, CDER; 2009. Available at: <http://www.fda.gov/downloads/Drugs/Guidances/UCM193282.pdf>. Published December 2009. Accessed September 1, 2015.
- American Educational Research Association, the American Psychological Association, National Council on Measurement in Education. Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association; 2014.
- Muthén BO, Muthén LK. Mplus: statistical analysis with latent variables, version 4.1 [computer program], Los Angeles: Muthén & Muthén; 2006.
- Jöreskog KG, Sörbom D. LISREL (version 8.71) [computer program], Chicago: Scientific Software International; 2004.
- Bentler PM. Comparative fit indexes in structural models. *Psychol Bull* 1990;107:238–246.
- Browne MW, Cudeck R. Alternative ways of assessing model fit. In: Bollen KA, Long JS, eds. *Testing Structural Equation Models*. Newbury Park: Sage Publications; 1993;136–62.
- Hu LT, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct Equ Modeling* 1999;6:1–55.
- Samejima F. Estimation of Latent Ability Using a Response Pattern of Graded Scores (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society; 1969.
- Multilog (version 7) [computer program]. Lincolnwood, IL: Scientific Software International; 2003.
- Samejima F. Estimation of Latent Ability Using a Response Pattern of Graded Scores (Psychometric Monograph, No. 17). Richmond, VA: Psychometric Society; 1969.
- Kincaid JP, Fishburne RP, Jr, Rogers RL, Chissom BS. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel, Springfield, VA: National Technical Training Service, U.S. Department of Commerce; 1975. Available at: <http://www.dtic.mil/dtic/tr/fulltext/u2/a006655.pdf>. Accessed September 1, 2015.
- Wirth RJ, Edwards MC. Item factor analysis: current approaches and future directions. *Psychol Methods* 2007;12:58–79.

**R. J. Wirth**

*Vector Psychometric Group, LLC  
847 Emily Lane  
Chapel Hill, NC 27516  
e-mail: [rjwirth@vpgcentral.com](mailto:rjwirth@vpgcentral.com)*