The ROYAL COLLEGE of OPHTHALMOLOGISTS

ARTICLE

Check for updates

The watery eye quality of life (WEQOL) questionnaire: a patient-reported outcome measure for surgically amenable epiphora

Christopher B. Schulz[1,2]✉, Paul Rainsbury[1,3], Jeremy J. Hoffman[4,5], Laura Ah-Kye[4], Elizabeth Yang[4], Raman Malhotra[2], Simon Rogers[1], Peter Fayers[6] and Tessa Fayers[4]

OBJECTIVE OR PURPOSE: To develop and test a patient-reported outcome measure for assessing health-related quality of life (HRQOL) in surgically amenable epiphora.
DESIGN: Questionnaire development and validation study.
PARTICIPANTS: 201 patients with a cause of epiphora amenable to surgical intervention, recruited across three independent centres.
METHODS, INTERVENTION OR TESTING: The watery eye quality of life (WEQOL) questionnaire was developed and refined according to defined psychometric standards. Both surgical and non-surgical participants completed WEQOL at baseline and follow-up (>3 months), along with the Lacrimal Symptom Questionnaire (Lac-Q), RAND Short Form Health Survey (SF-36) and Glasgow Benefit Inventory (GBI). Convergent validity of WEQOL was evaluated according to correlation ($R > 0.40$) with each of these additional tests. Responsiveness of WEQOL to intervention was evaluated according to patient-reported success. Test-retest reliability was assessed by the Bland–Altman method and intraclass correlation (ICC) in a subset of 64 participants at baseline.
MAIN OUTCOME MEASURES: WEQOL construct validity, responsiveness and test-retest reliability.
RESULTS: WEQOL was moderately correlated ($R > 0.4$) with the Lac-Q and several subscales of the SF-36 (physical role limitation, social, emotional role limitation and emotional well-being). A stronger correlation was found between the change in WEQOL at follow-up and GBI ($R = 0.61$). An appropriate graded response was found with a significant change in WEQOL score being observed in patients reporting successful (−28%, $p < 0.0001$) and partially successful surgery (−6%, $p = 0.04$), but not in those reporting unsuccessful surgery (+2%, $p = 0.9$). High test-retest reliability was observed (ICC = 0.93).
CONCLUSIONS: The WEQOL questionnaire has been developed systematically according to modern psychometric standards and has been designed to evaluate the quality of life in patients with epiphora that is of a surgically amenable cause. In this study, it has demonstrated appropriate test-retest reliability, responsiveness and construct validity.

Eye (2022) 36:1468–1475; https://doi.org/10.1038/s41433-021-01674-z

INTRODUCTION

Epiphora (watery eye) is a common complaint amongst patients attending eye clinics. The underlying aetiology is widely variable and is often multifactorial [1–3]. Potential contributing factors include tear outflow obstruction, eyelid laxity or malposition, lacrimal hypersecretion and tear film instability. Most cases are at least in some part amenable to surgical correction [1]. Despite the variation in aetiology and potential for surgical management in the majority of relevant cases, all patients are unified by one overarching characteristic: that their health-related quality of life (HRQOL) is lessened by some degree because of their watery eye [4]. For some patients, any such impact on HRQOL might be minimal, while for others they may be far more debilitating. Indeed, the visual disability associated with epiphora has been compared with that of cataract [5, 6]. With the exception of those rarer cases where the priority is for life- or sight-saving outcomes, it is the degree of this impact on HRQOL that we primarily wish to address when a patient is offered surgery. It is this same outcome that ought to be measured when evaluating interventional success or when comparing outcomes of one intervention against another. Quality of life in patients with epiphora does not have a clear, precise and universally agreed definition that is directly or reliably measurable. Despite a range of questionnaires for patient-reported outcome measures (PROMs) being used widely in the related literature, none have yet been proven to be both psychometrically robust and clinically meaningful for measuring the impact of epiphora on quality of life [2]. The aim of this work was to develop and test a new instrument, the 'Watery Eye Quality of Life (WEQOL)'

[1]Department of Ophthalmology, Portsmouth Hospitals University NHS Trust, Cosham, Portsmouth PO6 3LY, UK. [2]Corneoplastic Unit, Queen Victoria Hospital NHS Foundation Trust, East Grinstead RH19 3DZ, UK. [3]Department of Ophthalmology, University Hospitals Plymouth NHS Trust, Plymouth PL6 8DH, UK. [4]Western Eye Hospital, Imperial College Healthcare NHS Trust, London NW1 5QH, UK. [5]International Centre for Eye Health, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK. [6]Institute of Applied Health Sciences, University of Aberdeen, Aberdeen AB25 2ZD, UK. ✉email: chrisschulz@doctors.org.uk

Received: 28 February 2021 Revised: 13 June 2021 Accepted: 25 June 2021
Published online: 7 July 2021

questionnaire, in accordance with modern psychometric standards, including Item Response Theory (IRT) [7, 8].

METHODS

Definition of target population and instrument purpose

IRT is a framework for designing, analysing and scoring questionnaires to measure variables or 'latent traits' that cannot be reliably or directly measured. IRT is a probabilistic model as opposed to classical test theory (CTT) which is founded on averages and correlations. Thus, in IRT the focus is on estimating, for every level of the latent trait, the probability that a respondent has endorsed a particular response to each question on the instrument. In contrast, in CTT the emphasis is on modelling the average scores reported by each respondent. Equivalently, CTT is similar to linear regression modelling whereas IRT can be considered related to logistic regression modelling. The intended latent trait to be assessed by WEQOL was defined as HRQOL in patients with surgically amenable epiphora. In contrast with most instruments previously reviewed in the literature [2], the intention was to consider not only the symptomatic and functional impact of epiphora, but also its effect on other relevant domains of overall quality of life. The target population was defined as any adult patient with epiphora of a surgically amenable cause. The specification of 'surgically amenable' is important so as to develop a questionnaire that is appropriate and sensitive for use in interventional studies.

WEQOL development

A fundamental aspect to the validity of any HRQOL measure is to ensure that those issues that are relevant to patients are qualitatively explored and adequately covered by the instrument being developed. Patients with potentially surgically amenable epiphora were recruited from outpatient oculoplastic clinics to participate in a total of three focus groups of 6–7 patients. The baseline characteristics are presented in Supplementary Material 1. These focus groups were facilitated by members of the research team (CBS, PR) to help identify common themes relevant to the impact of epiphora on an individual's QOL. The discussions were guided (but not limited) by constructs identified from literature review and prior consultation with a panel of four oculoplastic consultants with experience in managing epiphora. The sessions were audio recorded, transcribed and coded into recurring themes, which were then grouped into overarching domains (Supplementary Material 2). With key domains ('physical', 'social' and 'mental') and sub-themes identified, potential questionnaire items were proposed by a panel of five clinicians with oculoplastic and lacrimal expertise. The relevance and clarity of each item was evaluated leading to the refinement of some items and the omission of some others. Adequate item coverage was ensured by developing a minimum of two items for each theme identified in Supplementary Material 2. The resulting 'question bank' was formatted into a pilot questionnaire, which was completed, and pilot tested in a cohort of 30 patients undergoing surgery to treat epiphora. They were asked to comment on the relevance to QOL, clarity and ambiguity of each item. Items with low endorsement or completion rates (<0.05) were discarded, along with those where the reported relevance was <0.10. Items with ambiguity rates >0.05 or clarity <0.05 were refined. The resulting items were formatted and presented as a self-administered questionnaire, the pre-test WEQOL (Supplementary Material 3).

WEQOL analysis, refinement and validity testing

Patient cohort and study timings. Three centres were involved in evaluating the WEQOL questionnaire over a 3-year period: ophthalmology departments in two teaching hospitals and a smaller specialized tertiary referral centre for cornea and oculoplastic disorders. Institutional Review Board/Ethics Committee approval was obtained and study procedures adhered to the tenets of the Declaration of Helsinki. All participants provided informed consent for study procedures. At each centre, patients with surgically amenable epiphora were recruited from oculoplastic specialist outpatient clinics and theatre lists and asked to complete the WEQOL questionnaire at baseline. Patients undergoing subsequent surgery were asked to complete the WEQOL questionnaire again a minimum of 3 months post-operatively (follow-up). Not all surgically amenable cases necessarily proceed to surgery. For those not undergoing subsequent surgery, patients completed a follow-up WEQOL questionnaire a minimum of 3 months following administration of the baseline questionnaire.

Endorsement, IRT analysis and questionnaire refinement. HRQOL instruments often assess multiple dimensions. For example, the SF-36 explores physical and mental health domains and addresses eight health concepts. No items should be fully dependent on another item as this equates to redundancy of that item. Furthermore, there should be at least some spread in the responses to each item (item endorsement) across the test population otherwise that item is not conferring any additional knowledge about the patient's degree of QOL impairment.

Borrowed from CTT, 'item fit' is a term that describes how well each questionnaire item fits with a specific latent trait. 'Discriminative ability' describes how well the responses to a given questionnaire item can discriminate between respondents with differing levels of QOL impairment. To optimize a questionnaire's performance and validity, it is important to refine its items after evaluation of all of these features. While CTT has guided the development of many PROMs in the past, modern computing has facilitated the use of IRT. IRT is preferred over CTT for analysing unidimensional scales, and especially in situations where the sample size is fewer than several thousands of patients.

All statistical analyses were conducted using R (version 3.6.2; R Core Team; Vienna, Austria). Paired (stacked) baseline and follow-up questionnaires proceeded to IRT analysis. Before fitting an IRT graded response model (https://weqol.shinyapps.io/multiple/), the assumptions of unidimensionality (that there is a single predominant latent trait) and local independence were evaluated using bifactor analysis with a minimum threshold for item-factor loading pre-defined as >0.4 and for residual correlations <0.25 [8]. The empirical (marginal) reliability of the graded response model can be computed according to latent trait estimates for each completed questionnaire and the associated standard errors. Item fit was evaluated using mean square infit and outfit statistics. The distribution of the latent trait amongst respondents and the discriminative ability (thresholds) for each item were evaluated by constructing a Wright map and item characteristic curves [8, 9]. Items with low endorsement rates (<0.05), poor infit or outfit (<0.6 or >1.4), or weak discriminative ability in the test population were either removed or their responses were recoded [10]. After any such changes, IRT analysis was repeated until the pre-defined criteria were met and an agreeable solution was found.

Traditional summary scoring (adding scored items) is commonplace in many PROMs due to its ease of use. Unfortunately, this assumes that each item is equally important, and that each response to a given item is on a continuous interval scale, neither of which are likely to be true for many questionnaires. IRT analysis can be used to manipulate the weighting of item scores so that the final estimate of the latent trait (QOL impairment) is measured on a continuous, linear scale. This can produce a final score that is more intuitive, more accurate and removes noise. As such, two potential methods of scoring the WEQOL questionnaire were proposed for comparison during validity and reliability testing. Firstly, a raw score was calculated by simple summation of the coded item responses. Secondly, IRT scoring was computed according to the fitted graded response model and transformed to a scale of 0–100. The decision to evaluate both was based on the relative simplicity and accessibility of a raw score (e.g. for everyday clinical use) vs. the linear scaling and the improved statistical robustness that an IRT score offers.

Reliability testing. In a subgroup of participants recruited at two of the centres, the questionnaire was further completed an additional time pre-operatively to evaluate inter-session test-retest reliability, with this second questionnaire being undertaken 2–4 weeks later. The intraclass correlation coefficient and Bland–Altmann analysis were used to evaluate test-retest reliability for both raw and IRT scoring patterns.

Validity testing. Construct validity of the WEQOL score was assessed according to several pre-defined hypotheses:

(1) WEQOL should show at least moderate positive correlation (Pearson's $R > 0.40$) with the Lacrimal Symptom Questionnaire (Lac-Q). This is a measure of symptom severity and disease-specific social impact previously developed in patients with nasolacrimal duct obstruction (NLDO) [11]. WEQOL and Lac-Q should not demonstrate collinearity ($R > 0.80$), which would suggest that the impact on QOL related to epiphora measured by WEQOL is already adequately evaluated by Lac-Q.

(2) WEQOL should show at least moderate positive correlation ($R > 0.40$) with validated measures of global HRQOL. For this study, the 36 item RAND Short Form Health Survey (SF-36) [12, 13] was used in a

Table 1. Baseline characteristics of study participants ($n = 201$).

	Number of participants (%)	Mean ± SD
Gender		
Male	84 (41.8)	
Female	117 (58.2)	
Age		64.8 ± 14.8
Unilateral	121 (60.2)	
Primary cause in unilateral cases:		
Eyelid disorder/lacrimal pump failure	32 (15.9)	
Punctal	24 (11.9)	
Canalicular	1 (0.5)	
Nasolacrimal duct obstruction		
Complete	30 (14.9)	
Partial	1 (0.5)	
Hypersecretion (e.g. gustatory)	16 (8.0)	
Multifactorial	14 (7.0)	
Bilateral	80 (39.8)	
Primary cause in bilateral cases:		
Bilateral eyelid disorder	21 (10.4)	
Bilateral outflow obstruction (i.e. punctal, canalicular, or NLDO)	34 (16.9)	
Mixed pathology	25 (12.4)	
Proceeded to surgery		
Yes	182 (90.5)	
No, medically unfit	2 (1.0)	
No, patient offered surgery but declined	9 (4.5)	
No, other reason specified (e.g. further investigations)	8 (4.0)	
Time to follow-up		
Surgical cases, $n = 182$ (weeks after surgery)		19.7 ± 5.9
Non-surgical cases, $n = 19$ (weeks after baseline)		19.3 ± 7.3

subgroup of patients at one participating centre and self-administered by patients at baseline. It comprises eight subscales (physical function, physical role limitation, pain, general health, energy, social, emotional role limitation and emotional well-being). Pearson's coefficient was used to evaluate correlation between the WEQOL score (both raw and IRT) and each of these subscales.

(3) WEQOL should be responsive to change. To evaluate this, patients who underwent surgery at two centres were asked to state whether they felt that their surgery was either 'successful', 'partially successful' or 'not successful'. In both 'successful' and 'partially successful' cases, an improvement in WEQOL scores should be detectable and this difference should be greater in the former compared with the latter. Hypothesis testing was conducted using paired Student's *t* tests to detect differences between baseline and follow-up WEQOL scores with alpha = 0.05. Responsiveness of SF-36 and Lac-Q was evaluated in the same way for comparison.

(4) The Glasgow Benefit Inventory (GBI) is a measure of improvement in health-related QOL, initially designed and validated for patients undergoing ENT procedures [14]. It is only administered post intervention and so is prone to recall bias. Despite this, it has been used with some success in several interventional studies of lacrimal disease. The GBI is scored negatively so that a greater negative score confers a greater improvement in QOL. It was hypothesized that the difference between paired baseline and follow-up WEQOL scores should be at least moderately negatively correlated (Pearson's $R < -0.40$) with The GBI was completed in a subgroup of participants at two centres at the follow-up visit.

(5) Not all cases that are surgically amenable will proceed to surgery. In cases that did not proceed to surgery, the reason for this was categorized as one of the following: 'medically unfit'; 'patient offered surgery but declined'; or 'other reasons'. It is reasonable to theorize that patients who are offered surgery but decline feel less of an impact on their QOL than those who elect to proceed to surgery. The unpaired Student's *t* test was used to test the hypothesis that baseline WEQOL scores in this subgroup were lower than in those patients that did proceed to surgery.

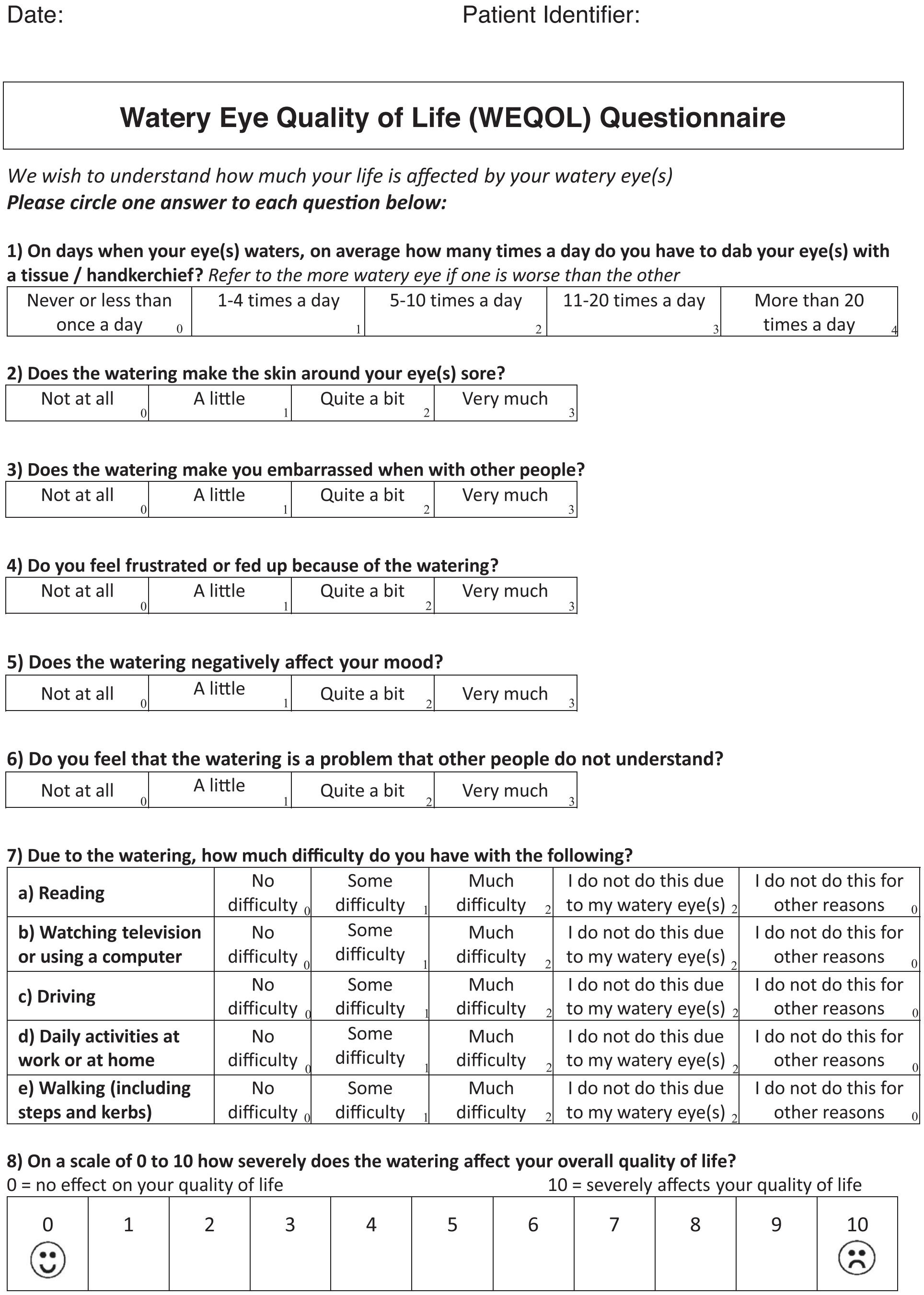
(6) It is likely (though not a given) that QOL is more negatively affected in more clinically severe disease states. Comparing disease severity across a wide and varied range of etiological causes is difficult. However, two tests were suggested. Firstly, it is plausible to consider that the impact of QOL might be greater in those patients that have bilateral epiphora vs. those with only one afflicted side. Secondly, it is considered that disease severity (and potentially the impact on QOL) might be greater in a subgroup of patients with complete NLDO, compared with those who have a partial obstruction. Based on the authors' clinical experience, the contentiousness of these two hypotheses was acknowledged. The sensitivity of WEQOL to detect each of these possible differences was tested by comparing scores (raw and IRT) using the unpaired Student's *t* test between groups.

RESULTS

In total, the pre-test WEQOL (Supplementary Material 3) was evaluated in 201 patients with surgically amenable epiphora. This comprised 104 patients in centre 1, 60 in centre 2 and 37 in centre 3. Patient characteristics are presented in Table 1.

Endorsement, IRT analysis and questionnaire refinement

All participants completed the baseline WEQOL questionnaire. Apart from 17 participants (8.5%) lost to follow-up, follow-up data were completed on the remaining 184 participants. In total, 385 completed questionnaires proceeded to initial analysis. The pre-test version of the questionnaire used is presented in Supplementary Material 3. There was adequate variation in item response with the exception of those questions pertaining to physical impact (items 9a–9e). In each of these questions, there were few respondents choosing the option 'I do not do this due to my watery eye(s)' (Supplementary Material 4). No items were negatively correlated or demonstrated collinearity >0.8. Question 1 ("In the past week, how many days have you had a watery eye?") was strongly correlated ($R > 0.7$) with questions 2 ("On days when your eye(s) waters, on average how many times a day do you have to dab your eye(s) with a tissue/handkerchief?") and also with question 3 ("When does watering occur?"). This suggests possible redundancy (Supplementary Material 4). Bifactor analysis indicated that all items adequately loaded onto a single unidimensional factor and residual correlations demonstrated adequate linear independence according to the pre-defined criteria (Supplementary Material 5). A graded response model was successfully fit to the stacked data with 58.6% of the raw variance observed in the latent trait accounted for by the IRT model. The empirical (marginal) reliability of this model was 0.93 with infit and outfit statistics for all items falling within the range of 0.60–1.40 (Supplementary Material 6). The model's Wright map (also presented in Supplementary Material 6) confirmed that the response 'I do not do this due to my watery eye(s)' for items 9a–9e had no discriminative ability in the tested population, with this response for 9c having particularly poor discriminative ability on characteristic curves. Based on these findings and the low endorsement of this response, the response was recoded to be of equal weighting to the previous response 'much difficulty'. The decision to recode this response option rather than remove it was based on maintaining



Date: Patient Identifier:

Watery Eye Quality of Life (WEQOL) Questionnaire

We wish to understand how much your life is affected by your watery eye(s)
Please circle one answer to each question below:

1) On days when your eye(s) waters, on average how many times a day do you have to dab your eye(s) with a tissue / handkerchief? *Refer to the more watery eye if one is worse than the other*

Never or less than once a day $_0$	1-4 times a day $_1$	5-10 times a day $_2$	11-20 times a day $_3$	More than 20 times a day $_4$

2) Does the watering make the skin around your eye(s) sore?

Not at all $_0$	A little $_1$	Quite a bit $_2$	Very much $_3$

3) Does the watering make you embarrassed when with other people?

Not at all $_0$	A little $_1$	Quite a bit $_2$	Very much $_3$

4) Do you feel frustrated or fed up because of the watering?

Not at all $_0$	A little $_1$	Quite a bit $_2$	Very much $_3$

5) Does the watering negatively affect your mood?

Not at all $_0$	A little $_1$	Quite a bit $_2$	Very much $_3$

6) Do you feel that the watering is a problem that other people do not understand?

Not at all $_0$	A little $_1$	Quite a bit $_2$	Very much $_3$

7) Due to the watering, how much difficulty do you have with the following?

a) Reading	No difficulty $_0$	Some difficulty $_1$	Much difficulty $_2$	I do not do this due to my watery eye(s) $_2$	I do not do this for other reasons $_0$
b) Watching television or using a computer	No difficulty $_0$	Some difficulty $_1$	Much difficulty $_2$	I do not do this due to my watery eye(s) $_2$	I do not do this for other reasons $_0$
c) Driving	No difficulty $_0$	Some difficulty $_1$	Much difficulty $_2$	I do not do this due to my watery eye(s) $_2$	I do not do this for other reasons $_0$
d) Daily activities at work or at home	No difficulty $_0$	Some difficulty $_1$	Much difficulty $_2$	I do not do this due to my watery eye(s) $_2$	I do not do this for other reasons $_0$
e) Walking (including steps and kerbs)	No difficulty $_0$	Some difficulty $_1$	Much difficulty $_2$	I do not do this due to my watery eye(s) $_2$	I do not do this for other reasons $_0$

8) On a scale of 0 to 10 how severely does the watering affect your overall quality of life?
0 = no effect on your quality of life 10 = severely affects your quality of life

0 ☺	1	2	3	4	5	6	7	8	9	10 ☹

Fig. 1 The revised WEQOL questionnaire. Subscripted numbers indicate summative value for each item response.

minimum ambiguity for the subsequent response 'I do not do this for other reasons'. The Wright map also confirmed the results of collinearity testing and that the thresholds and discriminative ability of questions 1, 2 and 3 were very similar, with 1 and 3 proving weakest on item characteristic curves (also available in supplementary Material 6). Based on this and the fact that item 2 is comparable with the well-established and widely recognized Munk score [15], questions 1 and 3 were removed from the questionnaire. Following this item refinement, the revised questionnaire evaluated is presented as Fig. 1. The resulting raw score (summation of coded responses) is on a scale of 0–39. The IRT logit score was transformed to a 0–100 scale. A graded response model was fit to the revised questionnaire with 56.9% of the raw variance in the latent trait explained by the model and an empirical reliability of 0.91. The model demonstrated appropriate item fit statistics and improved item characteristic curves (Supplementary Material 7).

Test-retest reliability

Retest WEQOL questionnaires were completed by 64 participants from two centres at a median of 15 days (range 14–30). The intraclass correlation coefficient was 0.93 using the IRT score and 0.95 using the raw score. A Bland–Altman plot for the test-retest agreement is presented for each of the scoring methods in Fig. 2, demonstrating similar limits of agreement taking into consideration the scale and distribution of responses.

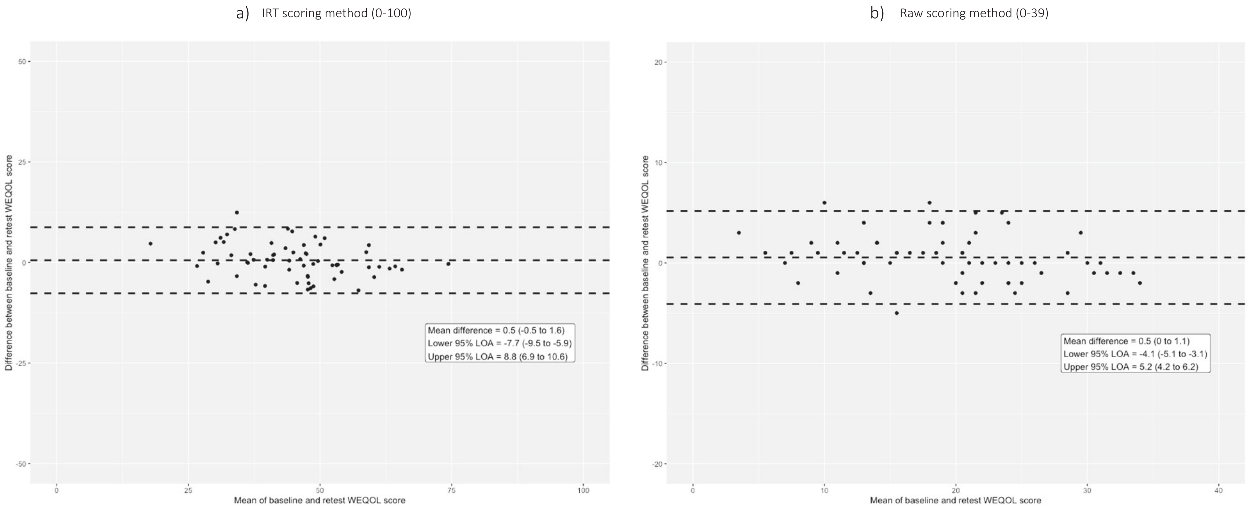


Fig. 2 Bland–Altman plot of agreement between baseline WEQOL and retest WEQOL in 64 participants with median 14 days between test and retest. IRT (**a**) and raw (**b**) scoring methods presented. Mean difference and Limits of agreement (LOA) marked by dashed lines and annotated with 95% confidence intervals.

Table 2. Convergent validity hypothesis tests correlation matrix.

	WEQOL score (IRT scoring method)	**WEQOL score (raw scoring method)**
	R	***R***
(1) SF-36 (*n* = 60)		
Physical function	−0.21	−0.17
Physical role limitation	−0.44	−0.44
Pain	−0.25	−0.22
General health	−0.24	−0.22
Energy	−0.37	−0.34
Social	−0.47	−0.48
Emotional role limitation	−0.46	−0.48
Emotional well-being	−0.42	−0.44
(2) Lac-Q		
Baseline (*n* = 103)	0.50	0.52
Follow-up (*n* = 96)	0.81	0.78
	Difference between baseline and follow-up WEQOL score (IRT scoring method)	**Difference between baseline and follow-up WEQOL score (Raw scoring method)**
	R	***R***
(3) GBI (*n* = 115)	−0.61	−0.56

Correlation coefficients (R) presented for (1) WEQOL vs. SF-36 subscales; (2) WEQOL score vs. Lac-Q; and (3) the change in WEQOL between baseline and follow-up vs. Glasgow Benefit Inventory (GBI) score. IRT (left) and Raw (right) scoring methods compared.

Construct validity tests

The 36-item RAND Short Form Health Survey (SF-36) was completed by 60 participants at one centre at baseline. Both WEQOL scoring methods demonstrated moderate correlation with several subscales of SF-36: physical role limitation, social, emotional role limitation and emotional well-being (Table 2). WEQOL was poorly correlated with the physical function, pain, general health or energy subscales. The Lac-Q was completed by 103 participants at a second unit at baseline, with 96 of those completing follow-up questionnaires (seven lost to follow-up). Lac-Q was moderately correlated with WEQOL scores at baseline, and more strongly correlated at follow-up (Table 2).

Regardless of the scoring method used, WEQOL demonstrated appropriate responsiveness to surgical intervention that was self-reported as ‘successful’ (*n* = 128) or ‘partially successful’ (*n* = 26) (Fig. 3). There was a mean reduction of 28.1 (95% C.I. 24.8–31.5; $p < 0.0001$) in IRT scores in successful surgery and 6.1 (0.2–12.1; $p = 0.04$) in partially successful surgery. The corresponding mean negative changes in raw scores were 14.5 (12.8–16.3; $p < 0.0001$) and 4.6 (1.1–8.1; $p = 0.01$). The difference in scores between baseline and follow-up was correlated with the GBI (Table 2). Lac-Q was found to be responsive to successful surgery (*n* = 50) (Fig. 3) but in this sample was not found to be sensitive enough to detect any meaningful change in patients whose surgery was reported as partially successful (*n* = 19).

There were nine study participants that were offered surgery but declined. The mean IRT WEQOL score for this group was 32.0, significantly lower than the 182 participants that elected to proceed with surgery (mean = 44.8; $p = 0.009$). There was no significant difference in WEQOL scores between unilateral

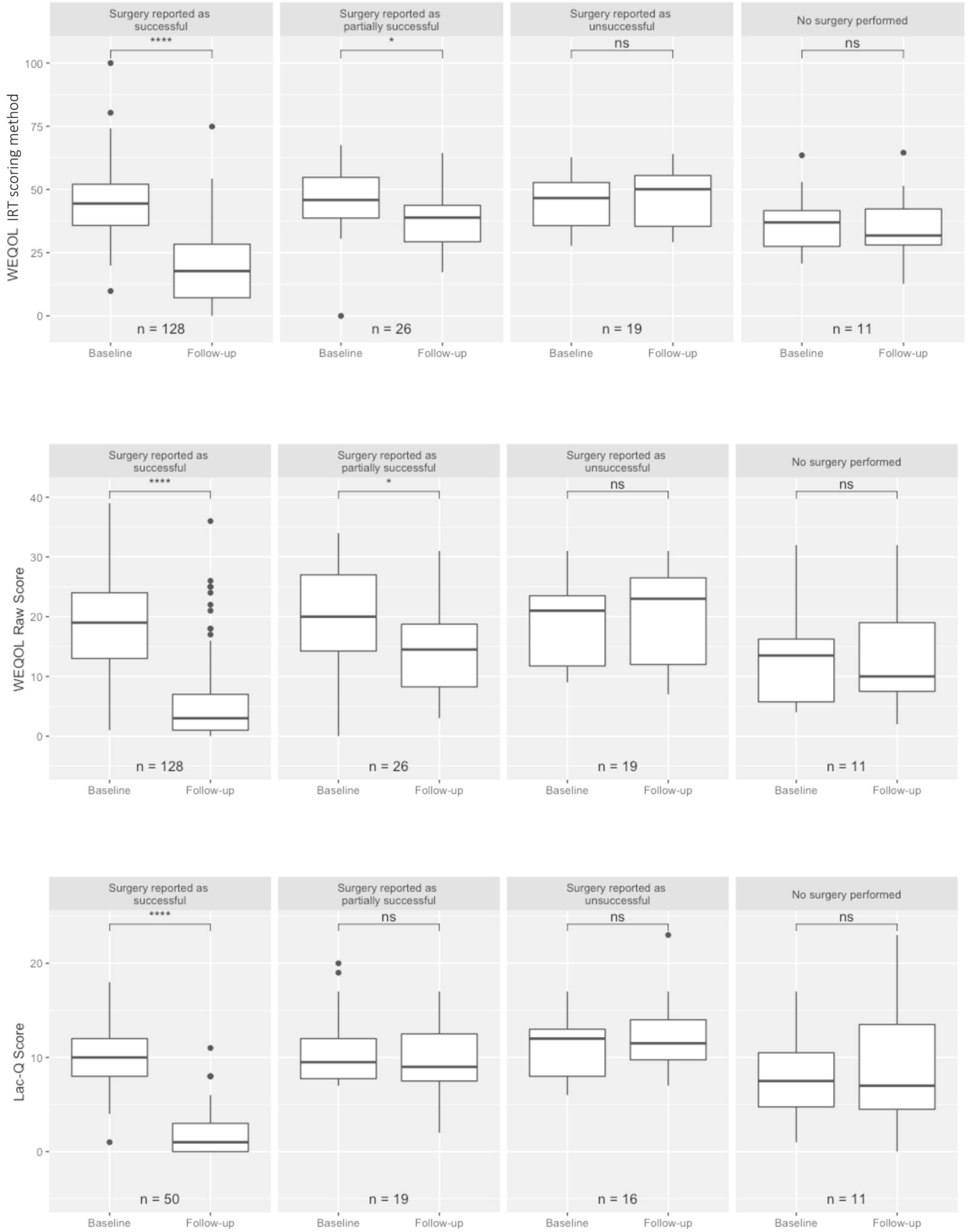


Fig. 3 Responsiveness to the surgical intervention of IRT WEQOL score (top panel), WEQOL raw score (middle panel) and Lac-Q (bottom panel). Participants grouped according to patient-reported success of surgery with non-surgical cases appended for comparison (right). Boxplots comprise median (thick line), 25th and 75th percentile (box), whiskers (1.5x interquartile range) and additional outliers (points). Paired *t*-test performed to compare differences in means: ****$p < 0.0001$; *$p < 0.05$; ns non-significant ($p \geq 0.05$).

($n = 121$) and bilateral disease ($n = 80$) states ($p = 0.4$). Nor was any difference detected in WEQOL scores between patients with complete ($n = 33$) and partial ($n = 11$) NLDO ($p = 0.4$).

DISCUSSION

The WEQOL questionnaire has been developed specifically to assess the quality of life in patients with surgically amenable epiphora. This enables us to evaluate the impact of surgery on QOL in these patients. The questionnaire was developed using a systematic framework for content development and refinement with thorough patient involvement from the outset, to ensure that it is valid and meaningful. For this target population and intended use, we are not aware of any other PROM which meets this standard [2]. The Lac-Q proposed by Mistry et al. [11] was developed for patients undergoing nasolacrimal surgery to quantify symptom severity. It is currently the only PROM designed for patients with epiphora that considers any recognized domain of health-related QOL: 'social impact.' No other potential domains of HRQOL are explicitly considered. The questionnaire items were identified by reviewing symptomatology and pathology in 100 patients referred with lacrimal obstruction, but it is unclear how this review was conducted or how the domains of 'social impact' were explored. The items of more global instruments measuring health-related QOL, such as the GBI [11] and SF-36 [10] have been developed in consultation with patients, but without being specific to those suffering with epiphora. To develop a valid and meaningful measure of QOL in patients with epiphora, it was our aim to explore and understand the themes that are relevant to the target population. In testing the WEQOL, it appears to measure a latent trait that is sufficiently unidimensional so that it can be adequately and meaningfully described by a global score. This global WEQOL score correlates moderately well to several subscales of the SF-36, a validated and well-established indicator of generic HRQOL: physical role limitation, social impact, emotional role limitation and emotional well-being. However, the ability of the present analysis to provide sufficient evidence that WEQOL adequately encompasses the pre-defined domains (physical, social, mental) is limited by the sample size of the study.

WEQOL also correlates moderately well with the Lac-Q as a measure of disease-specific epiphora-related symptom severity and social impact. The absence of a high correlation pre-operatively indicates that the two measures do not measure the same latent trait, perhaps due to the differences in how HRQOL was defined (Lac-Q focuses on symptom severity and social impact, whereas WEQOL was designed to include physical, social and mental aspects). The wider target population of WEQOL (all-cause surgically amenable epiphora) compared with Lac-Q (NLDO) is another likely root of this difference.

In measuring health-related QOL, global indicators are not always sensitive enough to detect change [8]. To demonstrate that the WEQOL questionnaire can be a useful and responsive outcome measure, as well as to further validate that the latent trait it measures behaves as expected, we have demonstrated its response to surgical intervention. Notably, it is highly responsive to 'successful' surgery and to a lesser degree surgery which is 'partially successful'. It remains stable in non-surgical cases or where surgery has been reported as 'unsuccessful' (Fig. 3).

In this study, the terms 'successful', 'partially successful' and 'unsuccessful' refer to patient-reported success. It is acknowledged that such a definition of success may be prone to recall bias and subjective interpretation [2]. In contrast, 'anatomical' success has often been reported in the literature [2]. However, it is our view that if our aim as lacrimal surgeons is to improve a patient's quality of life, it is the patient's perception of success that is of most relevance. To further support the validity of the WEQOL score, this study found that WEQOL scores are significantly lower in patients who elect not to proceed with surgery compared with those who do (acknowledging a limited sample size). With further evaluation, the WEQOL may not only be a useful outcome measure for patients with epiphora undergoing surgery but may also be useful to triage patients at the initial referral stage. In the final two proposed tests of construct validity, it was found that (1) WEQOL scores were similar in patients with bilateral disease compared with unilateral disease; (2) There was no difference in scores between patients with complete NLDO vs. partial NLDO. These findings are likely to be explained by one of two reasons. Either there truly is no difference between these subgroups in the latent trait measured by WEQOL or this sub-analysis was underpowered to detect a change that does exist. Even if the former is true, these two hypotheses of construct validity are arguably quite 'soft' theories. It is not known whether the impact on QOL is greater in patients with bilateral epiphora compared with their unilateral counterparts or if it is greater in complete NLDO compared with partial NLDO. Further study is required to better understand both theories.

In recent decades, IRT has become a mainstay of modern questionnaire development and item refinement because it overcomes many of the problematic assumptions of CTT. Using IRT analysis, we undertook systematic revision of the WEQOL questionnaire with item refinement. The revised version demonstrates promising psychometric properties, with well-fitting items and reliable discriminative ability. The reporting of psychometric testing, item refinement or questionnaire revision for pre-existing PROMs used to evaluate epiphora has been found to be lacking [2]. Regarding its reliability, the WEQOL questionnaire demonstrates promisingly high test-retest reliability and acceptable limits of inter-session agreement.

We evaluated both raw scoring and IRT scoring of the WEQOL questionnaire. The raw score comprises a simple summation of the coded responses in Fig. 1. It is simpler and easy to use by practitioners on-the-fly. IRT scoring relies on a more complex algorithm inherent to the fitted model. Though less accessible and user-friendly than the summation method, IRT scoring results in a linear scale without violating the assumptions that the items are of equal importance and that the responses are equally separated. In this study, both methods were shown to behave very similarly. As such, it is our recommendation that the raw scoring method can be appropriately used by clinicians or departments in routine practice. For clinical research where the WEQOL is subject to statistical analyses, the IRT scoring method is advised. To facilitate IRT scoring of the WEQOL, we have made two online resources publicly available: the first allows a score to be calculated for a single patient (https://weqol.shinyapps.io/individual/); the second calculate scores for a batch of multiple patients (https://weqol.shinyapps.io/multiple/).

CONCLUSION

The WEQOL questionnaire has been developed systematically in accordance with a standard modern framework for the development of PROMs. The latent trait it measures appears to be a valid marker of HRQOL in patients with a surgically amenable cause of epiphora. It correlates with global measures of HRQOL, and yet remains sensitive to changes in disease status. In most patients with epiphora, the primary goal of any intervention is to address its impact on quality of life. In such patients, the WEQOL questionnaire makes for an ideal candidate to measure this impact.

SUMMARY

What was known before

- Epiphora is a common presentation to the eye clinic and impacts patients quality of life to a variable degree. There is no universally agreed instrument for measuring health-related quality of life in patients with surgically amenable epiphora.

What this study adds

- The WEQOL questionnaire has been developed systematically in accordance with a modern standard for questionnaire development. WEQOL correlates with global measures of health-related quality of life, and yet remains sensitive to changes in disease status. In most patients with epiphora, the primary goal of any intervention is to address its impact on quality of life. In such patients, the WEQOL questionnaire makes for an ideal candidate to measure this impact.

REFERENCES

1. Woog JJ. The incidence of symptomatic acquired lacrimal outflow obstruction among residents of Olmsted County, Minnesota, 1976–2000 (an American Ophthalmological Society thesis). Trans Am Ophthalmol Soc. 2007;105:649–66.
2. Schulz CB, Kennedy A, Rogers S. A systematic review of patient-reported outcomes for surgically amenable epiphora. Ophthalmic Plast Reconstruct Surg. 2018;34:193–200.
3. Sibley D, Norris JH, Malhotra R. Management and outcomes of patients with epiphora referred to a specialist ophthalmic plastic unit. Clin Exp Ophthalmol. 2013;41:231–8.
4. Schulz C, Makuloluwe S, Rogers S. Issues affecting quality of life in patients with epiphora. Paper presented at the European Society of Ophthalmic Plastic and Reconstructive Surgeons Anual Meeting. Stockholm, Sweden; 2017.
5. Kafil-Hussain N, Khooshebah R. Clinical research, comparison of the subjective visual function in patients with epiphora and patients with second-eye cataract. Orbit. 2005;24:33–8.
6. Bohman E, Wyon M, Lundström M, Dafgård Kopp E. A comparison between patients with epiphora and cataract of the activity limitations they experience in daily life due to their visual disability. Acta Ophthalmol. 2018;96:77–80.
7. Pesudovs K, Burr JM, Harley C, Elliott DB. The development, assessment, and selection of questionnaires. Optom Vis Sci. 2007;84:663–74.
8. Fayers PM, Machin D. Quality of life: the assessment, analysis and reporting of patient-reported outcomes. Chichester: Wiley & Sons; 2016.
9. Wilson M, Bejar I, Scalise K, Templin J, Wiliam D, Irribarra DT. Perspectives on methodological issues. In: Assessment and teaching of 21st century skills. Vol. 19. 4th edn. Dordrecht: Springer; 2012. p. 67–141.
10. Wright BD, Linacre JM. Reasonable mean-square fit values. Rasch Meas Trans. 1994;8:370.
11. Mistry N, Rockley TJ, Reynolds T, Hopkins C. Development and validation of a symptom questionnaire for recording outcomes in adult lacrimal surgery. Rhinology. 2011;49:538–45.
12. Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36): I. Conceptual framework and item selection. Med Care. 1992;30:473–83.
13. Hays RD, Sherbourne CD, Mazel RM. The RAND 36-Item Health Survey 1.0. Health Econ. 1993;2:217–27.
14. Robinson K, Gatehouse S, Browning GG. Measuring patient benefit from otorhinolaryngological surgery and therapy. Ann Otol Rhinol Laryngol. 1996;105:415–22.
15. Munk PL, Lin DT, Morris DC. Epiphora: treatment by means of dacryocystoplasty with balloon dilation of the nasolacrimal drainage apparatus. Radiology. 1990;177:687–90.

AUTHOR CONTRIBUTIONS

CBS and TF were responsible for overall study conceptualisation, questionnaire development, methodology, analysis and manuscript preparation. PF supported the methodology, statistical analysis as well as manuscript review. PR, JJH, LA, EY, SR and RM supported the study with data collection, questionnaire refinement and manuscript review.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41433-021-01674-z.

Correspondence and requests for materials should be addressed to C.B.S.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.