

# The Oxford Finger Nail Appearance Score - a new scoring system for finger nail deformity.

Abhilash Jain<sup>1,2</sup>

Jamie Stokes<sup>1</sup>

Matthew D. Gardiner<sup>1,3</sup>

Jonathan Cook<sup>1</sup>

Amy Jones<sup>1</sup>

Cushla Cooper<sup>1</sup>

Beverly Shirkey<sup>1</sup>

Adam Sierakowski<sup>4</sup>

David Beard<sup>1</sup>

Aina V.H. Greig<sup>5</sup>

On behalf of the NINJA Collaborative \*

NINJA Collaborative \*

Sophie Dupré<sup>1</sup>, Raina Zarb Adami<sup>6</sup>, Benjamin Baker<sup>7</sup>, Malik Fleet<sup>8</sup>, Debbie Miles<sup>9</sup>, Rebecca Nicholas<sup>5</sup>, Alexi Nicola<sup>5</sup>, Agata Plonczak<sup>10</sup>, Aseel Sleiwah<sup>5</sup>, Georgina Williams<sup>2</sup>

**Corresponding author:** Matthew D. Gardiner

[Matthew.gardiner@kennedy.ox.ac.uk](mailto:Matthew.gardiner@kennedy.ox.ac.uk)

Department of Plastic and Reconstructive Surgery, Frimley Health NHS Foundation Trust, Slough, UK

## Author affiliations

1. Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK
2. Department of Plastic and Reconstructive Surgery, Imperial College Healthcare NHS Trust, London, UK
3. Department of Plastic and Reconstructive Surgery, Frimley Health NHS Foundation Trust, Slough, UK
4. St Andrew's Centre for Plastic Surgery and Burns, Mid Essex Hospital Services NHS Trust, Chelmsford, UK
5. Department of Plastic and Reconstructive Surgery, Guy's and St Thomas' NHS Foundation Trust, London, UK.
6. Department of Plastic Surgery, The Lister Hospital, Stevenage, UK
7. Department of Plastic Surgery, Manchester University Hospitals NHS Foundation Trust, Manchester, UK
8. Department of Plastic Surgery, Leeds General Infirmary, Leeds, UK
9. University of Essex, Colchester, UK
10. Department of Plastic Surgery, Countess of Chester Hospital NHS Foundation Trust, Chester, UK

## Summary

Finger nail deformity is common yet current methods used to define cosmetic appearance following trauma are mainly descriptive. In order to quantify cosmetic appearance of the finger nail we developed the Oxford Finger Nail Appearance Score using a three stage iterative process. The score has five cosmetic components marked as binary outcomes composed of nail shape, nail adherence, eponychial appearance, nail surface appearance and presence of a split. Twelve independent assessors scored 87 photographs of finger nails taken at a minimum four months following paediatric nail bed repair and compared them to the corresponding contralateral uninjured finger. Refinements in the scoring system resulted in an improvement in a weighted kappa statistic of 0.36 (95% CI:0.09,0.68) to 0.52 (95% CI: 0.42, 0.61). The Oxford Finger Nail Appearance Score is a user friendly and reliable scoring system which has application in a clinical trial setting.

### Key words

Finger Nail

Appearance

Oxford Score

Trauma

## INTRODUCTION

Nail deformity can be associated with systemic conditions and commonly occurs following infection and trauma<sup>1</sup>. The description and classification of finger nail deformity is subjective<sup>1,2</sup>. In order to quantify improvement in appearance following surgery Zook et al<sup>3</sup> developed a comparative scoring system. This was based on five domains (with between two and five subdomains) resulting in a summation score of major and minor variations. In 2017 we published the results of a pilot randomised control trial of paediatric nail bed injury (NINJA-P) in which 60 children were randomised to having the nail replaced or discarded following nail bed repair<sup>4</sup>. Cosmetic appearance was an outcome measure and the Zook score<sup>3</sup> was used to assess the cosmetic appearance of the injured nail four months after the injury/surgery had occurred.

The NINJA-P study identified agreement was poor between assessors for the Zook based score<sup>4</sup>. It was clear that the original Zook score developed over 35 years ago was not reliable for use in clinical trials to assess the outcome of nail appearance following trauma. As part of a large randomised control trial assessing cosmetic outcomes following nail bed repair<sup>5</sup> we modified the score to achieve greater consistency between assessors and added training to reduce inter-rater variability. The aim of the work was to design a score based upon the Zook score to improve the consistency and reliability of nail assessments in both clinical practice and clinical trial settings.

**METHODS**In the original NINJA-P study<sup>4</sup> two independent assessors used the Zook score<sup>3</sup> to assess 25 clinical photographs of the operated fingers of 25 children compared to the corresponding uninjured contralateral finger taken at the four-month time-point (Figure 1). Agreement between assessors was measured by way of Cohen’s kappa and percentage agreement on each of the score’s domains and the overall score. The overall score measure was created by taking a positive outcome for each component as a 1 and a negative outcome as a 0. The outcome for the components were added together to create a total score of between 0 and 5 (where 0 was the least optimal and 5 was the most optimal nail appearance). Percentage agreement in the nail shape and surface components was 36% and 48% respectively. Adherence, eponychium and split scored higher at 72%, 88% and 100% respectively. The agreement for the overall score was 40%, and a weighted kappa for the overall score was 0.36 (95% CI 0.09 to 0.68). In order to improve the reliability of the score we simplified the design of the Zook score by reducing the dimensionality of the score’s components (Figure 2). This was performed in two stages, as it was thought that the results obtained from the first set of modifications could be improved upon.

The reduction of dimensionality of the score involved making all five components of the Zook score binary by comparing the affected nail to the contralateral corresponding non-injured fingernail. The comparative components of the score (nail shape, eponychium and surface) had options reduced to either “identical to opposite” or “different to opposite” with regards to the contralateral finger. Adherence of the nail plate to the nail bed was modified to be scored as either “complete” or “incomplete”. The split component was left unchanged, as it was already scored as a binary measure. Assessors were also provided with training and reference sheets to assist with the completion of the scores. Examples were shown with characteristics to look for when scoring the nails (i.e. what a split nail might look like), and guidance was also provided in terms of how to complete the score correctly (i.e. pick only one option per component, fully complete all sections of the score).

A 'catch-all' style question was designed into the score, asking assessors if, in their opinion, enough of the nail plate had grown back to make a valid assessment of appearance. If this question was answered no, the assessor would not complete the score for that photograph. This was included to reduce potential bias which could arise from stunted growth caused by surgical site infections, as this has the potential to inhibit nail plate growth.

### *Photographic assessment*

This study was part of a larger randomised controlled trial<sup>5</sup> given ethical approval by the United Kingdom National Research Ethic Committee (18/SC/0024). All participants/parents/guardians gave consent for photographs to be used as part of this trial<sup>5</sup>. Ten assessors each completed the newly modified score on 62 photographs of the operated and uninjured contralateral corresponding finger used in the main NINJA study<sup>5</sup> at a minimum of 4 months following nail bed repair surgery. The assessors were made up of clinicians who scored the photographs independently of each other. Assessors were instructed to fully complete each of the five components. The overall component of the score was calculated by statisticians post-hoc taking the ideal appearance of each component as a 1 ("identical to opposite" for nail shape, eponychium and surface, "complete" for adherence, "absent" for split) and all the non-ideal appearances as 0. This meant that assessors effectively scored the photographs integer values between 0 (least optimal appearance) and 5 (most optimal appearance).

### *Statistical analysis*

Summaries of the assessors' scores were calculated for all 62 photographs. Numbers and percentages of the margin of disagreement between assessors were calculated. For example, the number of photographs that were given the same total score by all assessors and number of photographs scored

within 1 point for all assessors were calculated. The number of photographs with total agreement on both individual and all components were calculated along with percentages. The number and percentage of photographs which were deemed suitable for a valid assessment by all assessors were calculated.

As more than two assessors were now scoring the photographs, Cohen's kappa statistic<sup>6</sup> could not be used to calculate agreement in the same way as it was in the NINJA-P study<sup>4</sup>. Cohen's kappa and its weighted version can only be used in the case of two assessors. Instead, Fleiss' kappa statistic<sup>7</sup> was used to measure inter-rater reliability as it can be used as a measure of agreement in the case of more than two assessors.

Kappa for the overall score was weighted using a predefined weighting matrix, which is given as:

$$M_{Wgt} = \begin{pmatrix} 1 & 0.8 & 0 & 0 & 0 & 0 \\ 0.8 & 1 & 0.8 & 0 & 0 & 0 \\ 0 & 0.8 & 1 & 0.8 & 0 & 0 \\ 0 & 0 & 0.8 & 1 & 0.8 & 0 \\ 0 & 0 & 0 & 0.8 & 1 & 0.8 \\ 0 & 0 & 0 & 0 & 0.8 & 1 \end{pmatrix}$$

These weights were chosen to allow for some marginal disagreement between assessors. It was thought that a disagreement of 1 point in the overall score is reasonable considering the subjective nature of the cosmetic components, so reasonably high weights were given in these cases.

Percentage agreement and Fleiss' kappa were calculated for each of the individual component scores along with their associated 95% confidence intervals. For the total score, a weighted version of Fleiss' kappa was also calculated using  $M_{Wgt}$  as the weighting matrix, together with the associated 95% confidence interval.

All statistical analyses were performed in Stata v15.1. Fleiss' kappa values were calculated using the user-written kappaetc command<sup>8</sup>. **RESULTS**

The first round of modifications led to some improvement in the Zook score. However, the levels of percentage agreement and the kappa statistic for the overall score were relatively low. This prompted the second round of modifications described above and the results of which are reported in Table 1.

Of the 62 photographs, 58 were given a valid assessment by all 10 of the assessors. Three of the photographs were deemed suitable for valid assessment by nine assessors, whilst one of the photographs was only deemed suitable for assessment by six assessors. All assessors agreed on the total calculated score in 12 (19.4%) of the 62 photographs. These 12 universally agreed scores were given when assessors deemed the nail in the photograph to have the most optimal appearance (i.e. a score of 5). Twenty one (33.9%) photographs received a range of scores from assessors within 1-point of each other. Only one (1.6%) photograph received assessments with some scores differing by more than three points (i.e. one assessor scored the photograph a 0, whilst another assessor scored it a 4).

Percentage agreement and Fleiss' Kappa statistic (weighted and unweighted using predefined weighting matrix  $M_{Wgt}$ ) were calculated for the scores given to the photographs. Results calculated for all 62 photographs are shown in Table 2. Results for the 58 photographs with valid assessments from all assessors are shown in Table 3.

Individual component agreement ranged from 76.7% (95% CI: 71.3, 82.0) in nail shape to 97.8% (95% CI: 95.6, 100.0) in nail split when all photograph assessments are included in the analysis. Fleiss' kappa values ranged from 0.29 (95% CI: 0.19, 0.39) in nail shape to 0.61 (95% CI: 0.48, 0.74) in nail surface.

For photographs with valid scores from all assessors, the individual component agreement ranged from 77.3% (95% CI: 71.6, 82.9) in nail shape to 98.2% (95 CI: 96.1, 100.0) in nail split. Fleiss' kappa values for the individual components ranged from 0.29 (95% CI: 0.18, 0.39) in nail shape to 0.63 (95% CI: 0.22, 1.00) in nail split.

The overall score agreement was 60.2% (95% CI: 53.5, 67.0) when all 62 photographs were analysed, and 61.9% (95% CI: 54.9, 68.9) when only the 58 photographs with valid scores from all assessors were considered. Fleiss' kappa for all 62 photographs was 0.34 (95% CI: 0.27, 0.41) when unweighted and 0.51 (95% CI: 0.42, 0.59) when weighted with  $M_{Wgt}$ . Fleiss' kappa when analysing the 58 photographs with valid scores was 0.36 (95% CI: 0.28, 0.43) when unweighted and 0.52 (95% CI: 0.42, 0.61) when weighted with  $M_{Wgt}$ .

Results from the 58 photographs with valid scores from all assessors were finally considered, as missing data was minimal and had little effect on the agreement and kappa scores for both the individual components and the overall score. The redesign of the Zook score into the Oxford Fingernail Appearance Score shows clear improvement in the measure in terms of agreement. From the NINJA Pilot study<sup>4</sup>, all agreement scores for the individual components have shown a marked improvement, with the exception of nail split (Figure 3). However, this small reduction from 100% to 98.2% agreement is not concerning and is likely due to the fact that NINJA-P had only two assessors scoring 25 photographs, whereas the new assessment consisted of ten assessors assessing 62 (58 completely assessed) photographs, introducing a higher chance of disagreement.

## DISCUSSION

Nail deformity is commonly seen following infection and trauma yet despite this there are no validated scoring systems by which to quantify deformity. Current methods use descriptive terms for appearance of the nail. However, when determining outcomes of treatment these make comparison difficult. Zook et al<sup>3</sup> attempted to quantify finger nail appearance using a score utilising five important components of nail appearance, namely: nail shape, adherence, appearance of eponychium, nail surface appearance and the presence of a split. These categories were further subdivided into up to five potential outcomes and a sum of these variations used to describe the outcome as excellent, very good, good, fair or poor. In their original paper<sup>3</sup> there is no description of how the score was developed or validated. It was clear from the results in our pilot study of paediatric nail bed repair<sup>4</sup> that agreement of appearance using the Zook score was poor given the complexity of the score and the subjective nature of cosmetic appearance. As part of our large main randomised control trial looking at paediatric nail bed repair (NINJA) 451 children were randomised and nail appearance at a minimum of 4 months following surgery was assessed<sup>5</sup>. The Zook score, as originally described, was not fit for purpose to use in this trial and therefore we undertook work to improve the scoring system.

A scoring system should be easy to implement and have both validity and reliability. The Zook score was clearly too complicated and poorly reproducible. The NINJA clinical trial team of methodologists, senior clinicians and statisticians developed a more usable and reproducible finger nail appearance score over a series of step wise refinements to the original Zook score<sup>3</sup>. We simplified the scoring system based on results that were deemed clinically and patient relevant. Following any treatment the ideal outcome would be normal nail appearance and any deviation from this would be notable. However, whether the nail was for example “slightly rough” or “very rough” as described in the original Zook score was subjective and not relevant to patients as seen in that the majority of parent/patient responses to nail appearance using a visual analogue scale in NINJA-P suggested

normal appearance<sup>4</sup>. By making responses binary and by providing training and examples in the use of the Oxford score we demonstrated a marked improvement in agreement.

The kappa values from the NINJA pilot study and the newly modified score are not directly comparable, due to the former being a Cohen's kappa and the latter being Fleiss' kappa and the number of assessors/photographs differing between each analysis. However, the benchmarking scheme proposed by Landis and Koch<sup>9</sup> can be used to help interpret the kappa statistics and make an indirect comparison (Table 4).

The modifications made to the Oxford Finger Nail Appearance Score led to a beneficial effect on the weighted kappa statistic calculated for the overall score. The weighted kappa statistic from the original NINJA-P study was 0.36 (95% CI: 0.09, 0.68), which would indicate fair agreement according to the Landis and Koch interpretation. However, the new modified Oxford score gave a weighted kappa of 0.52 (95% CI: 0.42, 0.61), indicating moderate agreement between the assessors. Whilst still not as high as perhaps desired, this is likely due to the overall score being a composite of the five separate domains of the Oxford Finger Nail Appearance Score and the subjective nature of cosmesis. Whilst the levels of agreement for each individual component are relatively high as seen previously, disagreement in any of these components between assessors is likely to be compounded when assessing the agreement and kappa statistic of the overall score, which could have led to the lower kappa statistic.

The Oxford Finger Nail Appearance Score provides a more user friendly and reliable means of assessing the cosmetic appearance of finger nails in a trial situation which has implications for its use to judge response to treatment not just following surgery but following infection or other systemic conditions

affecting nail appearance. While the score has been developed for finger nail appearance there is no reason why it is not applicable to the toe nail as well.

### **Funding**

AJ, MDG, CC, AS, JC, DB, AVHG obtained grant funding for this project. This paper presents independent research funded by the National Institute for Health Research (NIHR) under its Research for Patient Benefit (RfPB) Programme (Grant Reference Number PB-PG-1215-20041) and was supported by the NIHR Biomedical Research Centre. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

### **Competing interests**

None to declare.

### **Ethics approval**

The National Research Ethic Committee approved this study on 2nd February 2018 (18/SC/0024).

## REFERENCES

1. Zook EG, Russell RC .Reconstruction of a functional and esthetic nail. Hand Clin. 1990; 6; 59-68.
2. Rai A, Jha MK, Makhija LK, Bhattacharya S, Sethi N, Baranwal S. An algorithmic approach to posttraumatic nail deformities based on anatomical classification. J Plast Reconstruct Aesthet Surg 2014; 67; 540-7.
3. Zook EG, Guy RJ, Russell RC. A study of nail bed injuries: causes, treatment, and prognosis. J Hand Surg Am 1984; 9; 247–252.

4. Greig A, Gardiner M, Sierakowski A et al. Randomized feasibility trial of replacing or discarding the nail plate after nail-bed repair in children. *Br J Surg* 2017; 104; 1634-1639.
  
5. Jain A, Jones A, Gardiner MD et al. NINJA trial: should the nail plate be replaced or discarded after nail bed repair in children? Protocol for a multicentre randomised control trial. *BMJ Open* 2019; 9; e031552.
  
6. Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 1960; 20; 37-46.
  
7. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 1971; 76; 378.
  
8. Klein D. Implementing a general framework for assessing interrater agreement in Stata. *The Stata Journal* 2018; 18; 871-901.
  
9. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 1; 159-74.

## **FIGURE LEGENDS**

Figure 1. Photographs of previously operated nail bed and contralateral non-operated nail bed at 4 months following surgery. Photograph A demonstrates near identical appearance of the two nails whereas photograph B demonstrates incomplete adherence of the injured digit nail (arrow) compared to the uninjured digit.

Figure 2. Oxford Finger Nail Appearance Score as modified from the original Zook et al paper<sup>3</sup>.

Figure 3 - Comparison of percentage agreement scores between pilot study assessments<sup>4</sup> and redesigned score.

Table 1: Characteristic summaries of photograph scores

	n (%)
<b>Total agreement* on valid assessment suitability</b>	
Yes	58 (93.6)
<b>Range of total scores between all assessors</b>	
0-point difference**	12 (19.4)
1-point difference**	21 (33.9)
2-point difference**	18 (29.0)
3-point difference**	10 (16.1)
>3-point difference**	1 (1.6)
<b>Photographs with total agreement* by component</b>	
Shape	23 (37.1)

Adherence	27 (43.5)
Eponychium	49 (79.0)
Surface	37 (59.7)
Split	55 (88.7)

*\*All assessors agreed on the component score given*

*\*\*Between min and max scores given by each assessor per photograph*

Table 2: Agreement and kappa scores calculated for the Oxford Finger Nail Appearance Score after the second round of modifications. Results shown for all 62 photographs.

<b>Component</b>	<b>% Agreement</b>	<b>95% CI</b>	<b>Fleiss' Kappa</b>	<b>95% CI</b>
Nail Shape	76.7	[71.3, 82.0]	0.29	[0.19, 0.39]
Nail Adherence	80.4	[75.4, 85.5]	0.45	[0.33, 0.57]
Eponychium	93.9	[90.5, 97.4]	0.26	[0.04, 0.47]
Nail Surface	86.7	[82.0, 91.4]	0.61	[0.48, 0.74]
Split	97.8	[95.6, 100.0]	0.56	[0.12, 0.99]
Overall (unweighted)	60.2	[53.5, 67.0]	0.34	[0.27, 0.41]
Overall (weighted)	-	-	0.51	[0.42, 0.59]

*Confidence intervals are bounded at 0 and 1*

*Agreement and kappa are unweighted unless specified*

Table 3: Agreement and kappa scores calculated for the Oxford Finger Nail Appearance Score after the second round of modifications. Results shown for 58 photographs receiving valid score from all assessors.

<b>Component</b>	<b>% Agreement</b>	<b>95% CI</b>	<b>Fleiss' Kappa</b>	<b>95% CI</b>
Nail Shape	77.3	[71.6, 82.9]	0.29	[0.18, 0.39]
Nail Adherence	81.9	[76.8, 87.0]	0.48	[0.36, 0.60]
Eponychium	94.7	[91.9, 98.3]	0.30	[0.08, 0.52]
Nail Surface	87.4	[82.6, 92.2]	0.61	[0.47, 0.75]
Split	98.2	[96.1, 100.0]	0.63	[0.22, 1.00]
Overall (unweighted)	61.9	[54.9, 68.9]	0.36	[0.28, 0.43]
Overall (weighted)	-	-	0.52	[0.42, 0.61]

*Confidence intervals are bounded at 0 and 1*

*Agreement and kappa are unweighted unless specified*

Table 4: Landis and Koch (1977) interpretation of kappa values

<b>Value of Kappa</b>	<b>Interpretation</b>
< 0.2	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement