# Article ▸ The Influence of Practitioner Experience on Inter-Observer Agreement in Performance of Cup-To-Disc Ratio Measurements

Brett O'Connor, BS, Class of 2016, Southern College of Optometry
Memphis, Tennessee

## ABSTRACT

**Background:** The cup-to-disc (C/D) ratio is a commonly used measurement for assessing the optic nerve requiring the use of monocular and binocular visual processing. Inter-observer agreement is known to be poorer than intra-observer agreement with regard to the precision of C/D ratio measurements, but studies evaluating reproducibility in live subjects and across different experience groups are scarce. While previous studies use stereoscopic photographs and expert observers to estimate reproducibility, this study evaluated inter-observer variability in live subjects among both students and practicing clinicians.

**Methods:** Eleven observers from each of four experience groups measured the right optic nerves of three subjects via slit lamp fundoscopy. For each experience group, limits of reproducibility were constructed for vertical and horizontal C/D ratio measurements by subject and analyzed for significant differences. Furthermore, data were analyzed in a categorical fashion in which differences between corresponding intra-class measurements were classified as falling within 0.05, 0.10, 0.15, or more than 0.15 C/D ratio units of one another and analyzed for trends.

**Results:** Limits of reproducibility ranged from 0.07 to 0.32 C/D ratio units, with no significant trend toward superior reproducibility in the groups with more clinical experience. Furthermore, a significant trend indicating improvement in agreement corresponding with increased experience was not exhibited in the higher experience levels via the categorical approach.

**Conclusions:** Reproducibility of C/D ratio measurements between observers is relatively broad and does not appear to correlate strongly with practitioner experience. When possible, multi-clinician practices should attempt to have the same clinician monitor C/D ratios in order to eliminate inter-observer variability. Poor binocularity and/or visual processing skills of some clinicians may contribute to these findings.

**Keywords:** cup-to-disc ratio, intra-observer agreement, reproducibility, stereopsis, visual performance

## Introduction

Measurement of the cup-to-disc (C/D) ratio and neuroretinal rim tissue is routinely used during examination of the posterior segment in the diagnosis and management of glaucoma and other optic neuropathies.[1] While advances in technologies such as optical coherence tomography and automated perimetry provide clinicians additional data for clinical decision-making, the C/D ratio remains a simple and universal technique for evaluating the optic nerve head.

Visual performance of the practitioner plays a critical role in the precision of his or her C/D ratio measurements. Using monocular depth cues (e.g., color change and defocus blur) in addition to binocular cues (e.g., stereopsis) enables the clinician to estimate the relative proportion of a patient's intraocular optic nerve that comprises the cup. The C/D ratio measurement requires high-level visual information processing on the part of the observer; as such, the visual skills of the practitioner play an important role in influencing inter-observer agreement between measurements. Since visual performance varies among clinicians, it is important to determine the reproducibility of C/D ratio measurements in order to identify whether a significant increase (cupping) has occurred between visits.

It is well-established that intra-observer agreement (repeatability) is superior to inter-observer agreement (reproducibility) with regard to measurement of the C/D ratio.[1-7] In a clinical environment in which the same practitioner routinely evaluates patients over time, inter-observer variability is effectively eliminated. However, in academic institutions and multi-clinician practices where patients are often examined by different practitioners over time, inter-observer variability must be considered as a factor affecting measurement precision.

Research evaluating repeatability and reproducibility of C/D ratio measurements has historically used "expert observers" and optic nerve photographs (often stereoscopic) to estimate agreement.[1-10] One major drawback to using optic nerve photographs is the reduction or elimination of many of the depth cues available through biomicroscopy. Practically speaking, clinicians estimate C/D ratios in live patients via fundoscopy, therefore the use of photographs in clinical trials may not provide a realistic estimation of variability. Additionally, the use of expert observers is problematic due to wide variation in the definition of an expert observer across the literature. Since a large proportion of glaucoma patients are managed by eye care professionals who may not fit the criteria of an expert observer, these studies may not provide a realistic approximation of inter-observer variability across the general population of practitioners.

Some evidence suggests that while the repeatability of C/D ratio measurements is not strongly correlated with clinical experience, reproducibility does tend to show a correlation with experience level of the practitioner.[7] However, studies evaluating inter-observer agreement of C/D ratio measurements in live subjects (rather than photographs) are lacking. This study aims to address this gap in the literature by determining the degree to which clinical experience influences inter-observer variability of C/D ratio measurements in realistic conditions.

## Methods

Study participants were recruited from the faculty and student body at Southern College of Optometry on a volunteer basis. Each participant provided informed consent, and this study was carried out in accordance with the Declaration of Helsinki. Three subjects (hereafter "Targets") were recruited to serve as mock patients to have their right optic nerves assessed. Targets were screened prior to study enrollment, and candidates with optic nerves displaying pathology or anomalous findings (e.g., optic nerve head drusen or disc malinsertion) were excluded.

Forty-four subjects (hereafter "Observers") were recruited from four experience groups based on clinical experience: second-year optometry students (Yr2), third-year optometry students (Yr3), fourth-year optometry students (Yr4), and practicing optometrists (OD), for a total of eleven Observers per Experience Group. In randomized fashion, Observers measured the C/D ratio of each Target's right optic nerve using Volk Digital High Mag lenses. Observers were permitted to modify beam width and

magnification as desired. Vertical and horizontal C/D ratios were gathered for each Target-Observer combination.

The C/D ratio data were evaluated in two ways. First, analyzing C/D ratio measurements as a continuous variable, limits of reproducibility were constructed for vertical and horizontal measurements for each Experience Group-Target combination. The reproducibility limit represents the value less than or equal to which the absolute difference between two test results obtained under reproducibility conditions may be expected to be with a probability of 95% (ISO 5725-1:1994). Reproducibility limits (LoR) were calculated as follows, where $\sigma_{repr}$ is the standard deviation of the measurements for a particular Experience Group-Target combination.

$$LoR = (1.96)(\sqrt{2})(\sigma_{repr})$$

A clinically significant difference in reproducibility between Experience Groups was defined as a difference of 0.05 or more, where the Experience Group with the smaller value was found to be clinically superior. Differences between Experience Groups of less than 0.05 were classified as indicative of clinical equivalence.

Second, the data were also analyzed using C/D ratio measurements as a continuous variable. Traditionally, Cohen's weighted kappa analysis has been utilized to provide a categorical estimation of reproducibility in C/D ratio studies.[1-9] The advantage to this methodology is that it enables greater weight to be assigned to closer measurements between observers (i.e., C/D ratios differing by 0.05 receive greater weight than those differing by 0.20). However, a universally accepted weight matrix has not been established in the literature. As such, significant variation in calculating Cohen's weighted kappa between trials makes it difficult to draw meaningful comparisons between studies.

In order to avoid this dilemma, a different statistical technique was utilized to categorically assess intra-observer agreement in this study. For each Target, vertical and horizontal C/D ratios were independently analyzed for all four Experience Groups. Each individual measurement (e.g., Yr2 Operator 1 – Target A – horizontal C/D) was compared to the corresponding measurements in that Experience Group (e.g., all other Yr2 Operators' horizontal measurements of Target A). Thus, each measurement was compared to ten other corresponding measurements. The difference between each pair of measurements (Yr2 Operator 1 vs. Yr2 Operator 2, Yr2 Operator 1 vs. Yr2 Operator 3, etc.) was classified as being within 0.05, within 0.10, within 0.15, or greater than 0.15 C/D ratio units. For each Target-Experience Group-Vertical/Horizontal grouping, the percentage of measurements falling within 0.05, 0.10, and 0.15 was calculated out of 110 comparisons (11 observers x 10 comparisons each).

## Results

### Continuous Analysis

Means, standard deviations, and limits of reproducibility for each Experience Group are given in Tables 1-3. The mean vertical and horizontal C/D ratios for each Target were within 0.05 C/D ratio units for each Experience Group, with the exception of Target A, in which Yr2 Observers significantly over-estimated both the vertical and horizontal C/D ratio compared to the other Experience Groups.

Limits of reproducibility for Target A ranged from 0.16 to 0.19 (all within 0.05). Thus, all Experience Groups were found to be clinically equivalent in terms of reproducibility for both vertical and horizontal measurements.

Limits of reproducibility for Target B ranged from 0.15 to 0.29. For the vertical measurement, the Yr3 value was clinically inferior to all other Experience Groups; Yr2, Yr4, and OD values were clinically equivalent to one another. For the horizontal measurement, the Yr3 value was clinically inferior to the OD value but clinically equivalent to the Yr2 and Yr4 values; the Yr2,

**Table 1: Target A – Continuous analysis**

| | Vertical | | Horizontal | |
|---|---|---|---|---|
| | Mean (SD) | LOR | Mean (SD) | LOR |
| **2nd years** | 0.2409 (0.0583) | 0.1618 | 0.2318 (0.0681) | 0.1887 |
| **3rd years** | 0.1273 (0.0607) | 0.1682 | 0.1409 (0.0701) | 0.1942 |
| **4th years** | 0.1364 (0.0636) | 0.1763 | 0.1364 (0.0636) | 0.1763 |
| **ODs** | 0.1364 (0.0595) | 0.1650 | 0.1364 (0.0595) | 0.1650 |

**Table 2: Target B – Continuous analysis**

| | Vertical | | Horizontal | |
|---|---|---|---|---|
| | Mean (SD) | LOR | Mean (SD) | LOR |
| **2nd years** | 0.6409 (0.0664) | 0.1841 | 0.6773 (0.0684) | 0.1897 |
| **3rd years** | 0.6773 (0.1057) | 0.2931 | 0.6955 (0.0789) | 0.2187 |
| **4th years** | 0.6864 (0.0636) | 0.1763 | 0.7000 (0.0632) | 0.1753 |
| **ODs** | 0.6409 (0.0769) | 0.2131 | 0.6864 (0.0552) | 0.1530 |

**Table 3: Target C – Continuous analysis**

| | Vertical | | Horizontal | |
|---|---|---|---|---|
| | Mean (SD) | LOR | Mean (SD) | LOR |
| **2nd years** | 0.2227 (0.0410) | 0.1137 | 0.2273 (0.0261) | 0.0724 |
| **3rd years** | 0.2364 (0.0393) | 0.1090 | 0.2455 (0.0522) | 0.1448 |
| **4th years** | 0.2045 (0.0650) | 0.1802 | 0.2136 (0.0636) | 0.1763 |
| **ODs** | 0.2227 (0.1034) | 0.2865 | 0.2364 (0.1164) | 0.3226 |

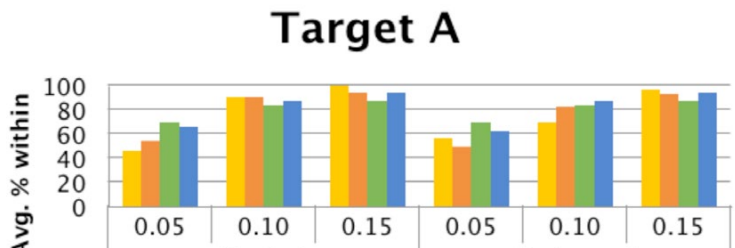
Yr4, and OD values were clinically equivalent to one another.

Limits of reproducibility for Target C ranged from 0.07 to 0.32. For the vertical measurement, Yr2 and Yr3 values were clinically equivalent, while Yr4 and OD values were clinically inferior to both Yr2 and Yr3 values; the OD value was furthermore clinically inferior to the Yr4 value. For the horizontal measurement, the Yr2 value was clinically superior to all other Experience Groups, Yr3 and Yr4 values were clinically equivalent to one another, and the OD value was also clinically inferior to Yr3 and Yr4.

## Categorical Analysis

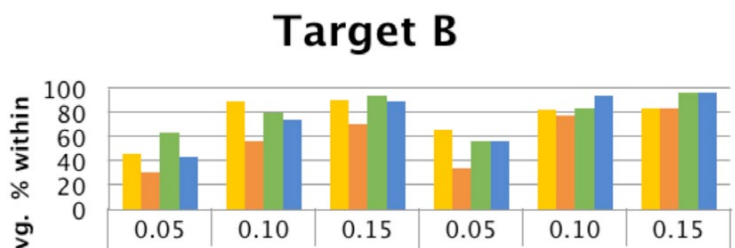
Figures/Tables 4-6 demonstrate the categorical analysis of the data, presented by Target and horizontal/vertical values. A value of 100 indicates that 100 percent of the corresponding observations fell within the given limit (0.05, 0.10, or 0.15 C/D ratio units).

**Table 4: Target A – Categorical analysis**



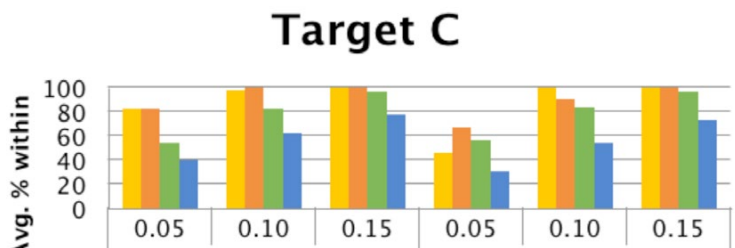

| | Vertical | | | Horizontal | | |
|---|---|---|---|---|---|---|
| | 0.05 | 0.10 | 0.15 | 0.05 | 0.10 | 0.15 |
| Yr2 | 45.5 | 90.9 | 100 | 56.4 | 69.1 | 96.3 |
| Yr3 | 54.5 | 90.9 | 94.5 | 49.1 | 81.8 | 92.7 |
| Yr4 | 69.1 | 83.6 | 87.3 | 69.1 | 83.6 | 87.3 |
| OD | 65.5 | 87.3 | 94.5 | 62.7 | 87.3 | 93.6 |

**Table 5: Target B – Categorical analysis**




| | Vertical | | | Horizontal | | |
|---|---|---|---|---|---|---|
| | 0.05 | 0.10 | 0.15 | 0.05 | 0.10 | 0.15 |
| Yr2 | 45.5 | 89.1 | 90.9 | 65.5 | 81.8 | 83.6 |
| Yr3 | 30.9 | 56.4 | 70.9 | 34.5 | 78.2 | 83.6 |
| Yr4 | 63.6 | 80 | 94.5 | 56.4 | 83.6 | 96.4 |
| OD | 43.6 | 74.5 | 89.1 | 56.4 | 94.5 | 96.4 |

**Table 6: Target C – Categorical analysis**




| | Vertical | | | Horizontal | | |
|---|---|---|---|---|---|---|
| | 0.05 | 0.10 | 0.15 | 0.05 | 0.10 | 0.15 |
| Yr2 | 81.8 | 98.2 | 100 | 45.5 | 100 | 100 |
| Yr3 | 81.8 | 100 | 100 | 67.3 | 90.9 | 100 |
| Yr4 | 54.5 | 81.8 | 96.4 | 56.4 | 83.6 | 96.4 |
| OD | 40 | 61.8 | 78.2 | 30.9 | 54.5 | 72.7 |

Significant, consistent improvement in agreement corresponding with increased experience was not exhibited in any of the three Targets, with Target C exhibiting poorer inter-observer agreement among the high experience levels.

## Discussion

Practically speaking, the absolute value of a C/D ratio is of little significance; rather, it is an increase in C/D ratio between visits which is most indicative of glaucomatous progression. The results of this study indicate that one can expect two observers to differ by 0.15 to 0.20 C/D ratio units roughly five percent of the time when viewing the same optic nerve, regardless of experience level. Since many clinicians regard an increase of 0.05 or 0.10 to be significant cupping, such a large limit of reproducibility makes it difficult to determine whether an increase in C/D ratio between visits is pathologic in nature, provided the measurements were made by different clinicians.

Over time, individual practitioners develop their own techniques for assessing C/D ratios. The individual visual performance of an eye care professional is influenced by training, visual memory, and variations in the use of various monocular and binocular cues. While these individual techniques may provide repeatable data for a single clinician, the results of this study indicate that comparisons between clinicians should be drawn with caution.

It is interesting to note that a significant trend toward superior reproducibility with more clinical experience was not noted with either the continuous or categorical approach. Moreover, in some scenarios the reverse was true. Thus, the results do not support the hypothesis that reproducibility among practitioners increases with clinical experience. It is possible that part of this phenomenon can be explained by more experienced clinicians placing more emphasis on other glaucomatous signs, such as the appearance of the neuroretinal rim tissue.

In light of this study's findings, it follows that multi-practitioner institutions should use objective measures of the optic nerve such as optical coherence tomography whenever possible, in addition to ensuring that glaucoma patients are monitored by the same clinician over time, thereby minimizing the effect of inter-observer error. Furthermore, C/D ratio grading systems such as those suggested by Lievens and colleagues[9] and Eiden and colleagues[11] may be warranted.

Future studies measuring stereoacuity of observers and correlating this data with reproducibility of C/D ratios would provide insight into the importance of binocularity in the assessment of the C/D ratio. Additionally, broader studies including a greater number of optic nerves would provide important information on whether the size of the optic nerve has an influence on the reproducibility of C/D ratio measurements.

*Special thanks to Dr. Patricia Cisarik for her assistance with this project.*

## References

1. Spalding JM, Litwak AB, Shufelt CL. Optic nerve evaluation among optometrists. Optom Vis Sci 2000 Sep;77(9):446-52. http://bit.ly/1i3hQPG
2. Tielsch JM, Katz J, Quigley HA, Miller NR, Sommer A. Intraobserver and interobserver agreement in measurement of optic disc characteristics. Ophthalmology 1988 Mar;95(3):350-6. http://bit.ly/1i3hXuL
3. Klein BE, Magli YL, Richie KA, Moss SE, et al. Quantitation of optic disc cupping. Ophthalmology 1985 Dec;92(12):1654-6. http://bit.ly/1FBpeY9
4. Varma R, Steinmann WC, Scott IU. Expert agreement in evaluating the optic disc for glaucoma. Ophthalmology 1992 Feb;99(2):215-21. http://bit.ly/1ES5yV1
5. Harper R, Radi N, Reeves BC, Fenerty C, et al. Agreement between ophthalmologists and optometrists in optic disc assessment: training implications for glaucoma co-management. Graefes Arch Clin Exp Ophthalmol 2001 Jun;239(5):342-50. http://bit.ly/1iC5gra
6. Abrams LS, Scott IU, Spaeth GL, Quigley HA, Varma R. Agreement among optometrists, ophthalmologists, and residents in evaluating the optic disc for glaucoma. Ophthalmology 1994 Oct;101(10):1662-7. http://bit.ly/1OifCsy
7. Hanson S, Krishnan SK, Phillips J. Observer experience and Cup:Disc ratio assessment. Optom Vis Sci 2001 Oct;78(10):701-5. http://bit.ly/1XSRhNX

8. Lichter PR. Variability of expert observers in evaluating the optic disc. Trans Am Ophthalmol Soc 1976;74:532-72. http://1.usa.gov/1VRtNH3

9. Lievens CW, Judd TA. Using the original Judd-Lievens C/D ratio grading card (JLC) to improve interobserver reliability. Optometric Education 2007 Winter;32(2):57-62. http://bit.ly/1ULdWay *abstract only

10. Feuer WJ, Parrish RK 2nd, Schiffman JC, Anderson DR, et al. The Ocular Hypertension Treatment Study: reproducibility of cup/disk ratio measurements over time at an optic disc reading center. Am J Ophthalmol 2002 Jan;133(1):19-28. http://bit.ly/1Oig222

11. Eiden SB, Cooper J, Olivares G, Horn D, London R. Interexaminer reliability of the optic cup to disc ratio assessment. Am J Optom Physiol Opt 1986 Sep;63(9):753-6. http://bit.ly/1OGhGZ6

*Correspondence regarding this article should be emailed to Brett O'Connor, BS, at boconnor@student.sco.edu.* 

O'Connor B. The influence of practitioner experience on inter-observer agreement in performance of cup-to-disc ratio measurements. Optom Vis Perf 2015;3(BSK):62-7.