OPO OPHTHALMIC & PHYSIOLOGICAL OPTICS
THE JOURNAL OF THE COLLEGE OF OPTOMETRISTS

Ophthalmic & Physiological Optics ISSN 0275-5408

INVITED REVIEW

# When to use the Bonferroni correction

Richard A. Armstrong

School of Life and Health Sciences, Aston University, Birmingham, UK

**Citation information:** Armstrong RA. When to use the Bonferroni correction. *Ophthalmic Physiol Opt* 2014; **34**: 502–508. doi: 10.1111/opo.12131

**Keywords:** Bonferroni correction, Clinical & Experimental Optometry, Ophthalmic & Physiological Optics, Optometry & Vision Science, statistical guidelines

*Correspondence:* Richard A. Armstrong
E-mail address: r.a.armstrong@aston.ac.uk

Received: 6 January 2014; Accepted: 13 March 2014; Published Online: 2 April 2014

## Abstract

*Purpose:* The Bonferroni correction adjusts probability ($p$) values because of the increased risk of a type I error when making multiple statistical tests. The routine use of this test has been criticised as deleterious to sound statistical judgment, testing the wrong hypothesis, and reducing the chance of a type I error but at the expense of a type II error; yet it remains popular in ophthalmic research. The purpose of this article was to survey the use of the Bonferroni correction in research articles published in three optometric journals, viz. *Ophthalmic & Physiological Optics*, *Optometry & Vision Science*, and *Clinical & Experimental Optometry*, and to provide advice to authors contemplating multiple testing.

*Recent findings:* Some authors ignored the problem of multiple testing while others used the method uncritically with no rationale or discussion. A variety of methods of correcting $p$ values were employed, the Bonferroni method being the single most popular. Bonferroni was used in a variety of circumstances, most commonly to correct the experiment-wise error rate when using multiple '*t*' tests or as a *post-hoc* procedure to correct the family-wise error rate following analysis of variance (ANOVA). Some studies quoted adjusted $p$ values incorrectly or gave an erroneous rationale.

*Summary:* Whether or not to use the Bonferroni correction depends on the circumstances of the study. It should not be used routinely and should be considered if: (1) a single test of the 'universal null hypothesis' ($H_o$) that *all* tests are not significant is required, (2) it is *imperative* to avoid a type I error, and (3) a large number of tests are carried out without preplanned hypotheses.

## Introduction

The Bonferroni correction, named after the Italian statistician Carlo Bonferroni (1892–1960), was based on a method proposed initially by Neyman and Pearson[1] to aid decisions in studies involving repetitive sampling. In modern research, however, the procedure is frequently used to adjust probability ($p$) values when making multiple statistical tests in any context and this usage is attributed largely to Dunn.[2] It has become a popular method and is widely used in various experimental contexts including: (1) comparing different groups at baseline, (2) studying the relationship between variables, and (3) examining more than one endpoint in clinical trials.[3,4] In addition, Bonferroni correction can be used to correct 'experiment-wise' and 'family-wise' error rates in multiple comparisons. Experiment-wise error correction is where a large number of independent tests are performed employing basic statistical procedures such as 'Students' ('*t*') or Pearson's correlation coefficient ('*r*') and all tests are included.[5] By contrast, family-wise error correction occurs when a smaller number of *related* group means are compared often after a *post-hoc* procedure following analysis of variance (ANOVA)[6–11] (also known as the Bonferroni *post-hoc* test).

The Bonferroni correction was proposed to circumvent the problem that as the number of tests increases, so does the likelihood of a type I error, i.e., concluding that a significant difference is present when it is not. Hence, if a null hypothesis ($H_o$) is true and $p \leq 0.05$ is used as the test criterion for all tests, a significant difference will be observed by chance one in 20 trials. If 20 tests are performed, and $H_o$ is true for all 20 tests, however, it can be shown that the chance of least one test being statistically significant is not $p = 0.05$ but $p = 0.64$.[12] In general, the error rate will be:

$$1 - (1 - \alpha)/T \quad (1)$$

where '$\alpha$' is the critical $p$ level and '$T$' is the number of tests performed. In practice, an adjusted significance level of $\alpha/T$ is used as an approximation to (1). Hence, the Bonferroni correction is applied to the $p$ values associated with each *individual* test to maintain the $\alpha$ level over *all* tests at 0.05.

Despite the widespread use of the Bonferroni method, there has been continuing controversy regarding its use. Hence, there are those who believe no correction should ever be made[13] and those who consider correction should be mandatory.[14,15] In addition, several criticisms have been made of the procedure, most notably by Perneger.[12] First, that the Bonferroni correction 'is at best unnecessary and at worst actually deleterious to sound statistical inference.' Second, that the method is often concerned with the wrong hypothesis and is in actuality a test of the 'universal' $H_o$, i.e., if 20 different comparisons were made on two groups, that the two groups were identical in *all* comparisons. Normally, this test would be of little relevance to researchers who wish to assess the statistical significance of individual tests. Third, that the interpretation of a single test depends on the number of other tests performed. Hence, it could be argued that the evidence provided by data is contained within that specific data set and therefore, the conclusion drawn should not be altered on the basis of the number of other tests performed. Fourth, the probability of a type I error cannot be decreased without increasing that of a type II error, such that real differences may not be detected. Hence, as the number of tests increases, the value of the adjusted $p$ that has to be exceeded to achieve statistical significance using the Bonferroni correction decreases markedly, lowering the power of a test. Fifth, there is the question of what constitutes the population of tests to which the correction should be applied, e.g., all tests in a report or a subset of them, tests performed but not included in the report, or tests from the same data included in other reports?[12] Given these criticisms and the popularity of the Bonferroni method among clinical researchers, a review of its use in optometric research would appear timely.

The purpose of this article is to review the use of the Bonferroni correction in ophthalmic research over the last 10 years and to provide some statistical advice for authors carrying out clinical studies which may involve the testing of multiple hypotheses. First, current practice in the use of Bonferroni and other types of correction is reviewed with reference to articles published in three optometric journals, viz. Ophthalmic & Physiological Optics (OPO), Optometry & Vision Sciences (OVS), and Clinical & Experimental Optometry (CXO). Second, statistical advice is given on the use of the Bonferroni correction in two statistical contexts: (1) correcting the experimental-wise error rate when making multiple tests involving a simple procedure such as '$t$' or '$r$' and (2) correcting the family-wise error rate following ANOVA.

## Methods

### Journals

All articles in which multiple statistical testing of data was employed and which were published in *OPO, OVS,* and *CXO* in the period 2003–2013 were reviewed. Two searches were made to investigate: (1) the frequency of correction of $p$ values by any available method (Search terms: 'multiple testing', '*post-hoc*' tests) and (2) the specific use of the Bonferroni adjustment (Search terms: 'Bonferroni correction', 'Bonferroni adjustment', 'Bonferroni *post-hoc* test'). Two questions were considered with reference to the articles examined: (1) did the article correct $p$ values to reduce the chance of a type I error using any of the available methods and provide a rationale for the method used (Search 1), and (2) did the study apply Bonferroni correctly and did it provide an appropriate rationale and/or discussion of its use (Search 2)?

### Data analysis

Differences in the distribution of frequencies between categories were compared among the three journals (totalled over years) using chi-square ($\chi^2$) contingency table tests.

## Results

The analysis of studies which included multiple statistical testing by any available method is shown in *Table 1*. Of 142 articles reviewed, 47 (33%) did not correct $p$ values for

**Table 1.** What proportion of studies in three optometry journals (*OPO = Ophthalmic & Physiological Optics, CXO = Clinical & Experimental Optometry, OVS = Optometry & Vision Science*) involving multiple statistical testing corrected probability ($p$) values to reduce the chance of a type I error and provided an appropriate rationale using any of the available methods?

| | Journal | | | |
|---|---|---|---|---|
| **Categories** | **OPO** | **CXO** | **OVS** | **Total** |
| Correction/Rationale | | | | |
| No correction | 17 | 9 | 21 | 47 |
| Correction, Rationale | 2 | 3 | 4 | 9 |
| Correction, no Rationale | 30 | 15 | 41 | 86 |
| Method | | | | |
| Bonferroni | 14 | 9 | 28 | 51 |
| Other | 35 | 18 | 38 | 91 |

Comparison of journals (Correction/Rationale: $\chi^2_4 = 1.58$ ($p = 0.81$).
Comparison of journals (Method: $\chi^2_2 = 2.44$ ($p = 0.30$).

multiple comparisons. Of the 95 (67%) of articles that did correct $p$ values, nine (9%) provided a clear rationale for its use, i.e., to avoid a type I error, while 86 (91%) provided no clear rationale or discussion. There were no differences in these proportions in the three journals ($\chi^2_4 = 1.58$, $p = 0.81$). The Bonferroni correction was specifically applied in 51 (36%) of articles, other types of correction such as the Bonferroni-Holm method, standard Abbott formula, the false discovery rate, the Hochberg method, or an alternative conservative *post-hoc* procedure, such as Scheffé's test, being used in the remainder. There were no significant differences in these proportions in the three journals studied ($\chi^2_2 = 2.44$, $p = 0.30$).

The proportion of studies in which either a clear rationale and/or the implications of the Bonferroni correction were given compared with those in which no such discussion was evident is shown in *Table 2*. Of 187 articles reviewed, 133 (71%) provided little or no discussion while 54 (29%) provided some rational and/or discussion of the method. Of the articles that provided some discussion, 36 (19%) considered its relevance in reducing a type I error, two (1%) discussed the possibility of a type II error, six (3%) discussed the relevance of the Bonferroni correction and decided not to adjust $p$ values, and eight (4%) gave an incorrect rationale for its use. Of the 187 articles reviewed, 72 (39%) explicitly stated the $p$ value used to judge statistical significance. Of these 34 (47%) quoted the adjusted $p$ value correctly while 38 (53%) continued to quote that $p = 0.05$ was used as test criterion even after adjustment. There were no significant differences in these proportions among the three journals (Discussion/Rational $\chi^2_{10} = 11.67$, $p = 0.31$; Quotation of $p$ value $\chi^2_2 = 4.80$, $p = 0.09$).

**Table 2.** What proportion of articles provide a rationale for the use of the Bonferroni correction and quoted adjusted probability ($p$) values correctly in three optometric journals (*OPO = Ophthalmic & Physiological Optics, CXO = Clinical & Experimental Optometry, OVS = Optometry & Vision Science)*

| | Journal | | | |
|---|---|---|---|---|
| **Categories** | **OPO** | **CXO** | **OVS** | **Total** |
| Dicussion/Rationale | | | | |
| No discussion | 50 | 32 | 51 | 133 |
| Discussion, type I error | 12 | 9 | 15 | 36 |
| Discussion, type II error | 2 | 0 | 0 | 2 |
| Discussion, no correction | 4 | 0 | 2 | 6 |
| Discussion with and without correction | 1 | 0 | 1 | 2 |
| Erroneous rationale | 6 | 0 | 2 | 8 |
| Adjusted $p$ | | | | |
| Correcly quoted | 7 | 6 | 21 | 34 |
| $p = 0.05$ quoted | 1 | 9 | 28 | 38 |

Comparison of journals: (Discussion/Rationale: $\chi^2_{10} = 11.67$ ($p = 0.31$). Quotation of adjusted $p$ ($\chi^2_2 = 4.80$ ($p = 0.09$).

## Discussion

A wide variety of clinical studies in optometry were reviewed and a range of current practice in correcting $p$ values when making multiple tests was identified. A variety of methods of correcting $p$ values were employed including Hochberg,[16] Greenhouse-Geisser,[17] false discovery rate,[18] the Abbott formula,[19] or 'guide to expression of uncertainty in measurement',[20] but the Bonferroni method was the single most popular. In addition, the greatest single use of these methods was as a *post-hoc* procedure following ANOVA[6–8] and its variants, including analysis of covariance (ANCOVA),[21] and multivariate ANOVA (MANOVA).[22]

Two main issues were identified in the studies reviewed. First, too many studies failed to address the problem of multiple testing, viz. the possibility of making a type I error and very few studies considered its corollary, i.e., the increasing risk of a type II error if a correction was applied. In addition, a few studies gave some consideration to the problem but then made the decision not to adjust $p$ values.[23–26] Some authors compared the results of both correcting and not correcting $p$ values[22] thus potentially complicating interpretation of the data. Second, when the problem of multiple testing was addressed by the application of the Bonferroni or an equivalent method, there were too many studies in which no rationale for its use or discussion of its effects was provided.

The Bonferroni correction itself was applied to a wide variety of statistical procedures, most frequently as a *post-hoc* test after ANOVA[6–8] or when multiple '$t$' tests[27–30] and Pearson's '$r$'[31–34] were employed. It was also used to correct non-parametric tests such as the Mann-Whitney test,[35] the Wilcoxon test,[36,37] the Kruskal-Wallis test,[38,39] chi-square ($\chi^2$) contingency table test,[40,41] and Fisher's $2 \times 2$ exact test.[42,43] It was less commonly used in studies involving regression and multiple regression,[29,44,45] or in studies involving the intraclass correlation coefficient (ICC),[46] and was rarely used to test goodness-of-fit of data to statistical models such as Rasch model or the normal distribution.[47,48] Applying a Bonferroni correction to a series of goodness of fit tests would not be recommended as reducing the chance of a type I error would increase a type II error, i.e., increasing the chance that some data sets would spuriously fit the model.

The majority of studies reviewed did not consider the relative risks of type I and type II errors. The relative importance attached to these two types of error will depend on the specific hypotheses tested and the likely consequences of making each type of error. For example, in an exploratory study, an ANOVA may have been carried out to determine which of a group of treatments or variables is likely to have a significant experimental effect, such effects being investigated in more detail. In this case, it would be

better to use a more liberal *post-hoc* test such as Fisher's 'protected least significant difference' (PLSD).[8] In this context, it is better not to miss a possible effect, i.e., to avoid a type II error and therefore not to use a Bonferroni correction. By contrast, if the objective is to be as certain as possible that a particular planned treatment comparison does have the desired effect then either the Bonferroni method or one of the more conservative *post-hoc* procedures such as Scheffé, would be more appropriate.[49,50] Hence, in any study involving multiple testing, authors should consider first, whether specific planned comparisons are envisaged (hypothesis testing) or whether unplanned exploratory testing is involved and second, consider the relative risks of type I and type II errors before making a decision of whether to adjust $p$ values.

A significant number of articles examined did not provide a rationale or any discussion of the method of correction used or its consequences. A further problem in those studies which made an adjustment was that many did not clearly quote the actual adjusted $p$ value in addition to the non-adjusted $p$ value and the name of the correction procedure. Some studies stated that a Bonferroni correction was made and that $p = 0.05$ remained the test criterion. This ambiguity could mean: (1) $p = 0.05$ was the original test criterion but was modified by a Bonferroni correction, (2) that after correction, the $p$ value remained at $p = 0.05$ over all tests, or (3) $p = 0.05$ continued to be used erroneously as at test criterion for the individual tests. A small number of studies used a Bonferroni correction but did not provide a correct rationale for doing so. Hence, a Bonferroni correction is not advisable in circumstances in which the variables under study are heavily inter-dependent[51] or to correct for unequal variances[52]; there are other methods available for taking these problems into account. In addition, applying the Bonferroni correction to a series of '$t$' tests to decide whether the data could be pooled for a subsequent ANOVA should be viewed with caution as the correction would encourage pooling by setting a particularly conservative $p$ value and thus increasing the chance of a type II error.[53] The Bonferroni correction has also been applied to multiple tests often on a single data set using different statistical procedures.[54–59] The problem with this approach is not necessarily the number of tests performed but that the various procedures are likely to be based on different statistical models, and may result in conflicting conclusions from the same data.

Given the criticisms of Pernerger[12], should the Bonferroni correction be used routinely? In the author's view, too many studies applied the correction uncritically, and since it is a conservative procedure, many 'real' effects may have gone undetected. Pernerger[12] describes a number of scenarios in which such a correction would be appropriate. First, as a test of the universal $H_o$ that all tests are not significant. For example, an investigator may wish to verify that a specific eye disease was not associated with the histocompatibility locus antigens (HLA), the occurrence of 20 specific antigens being measured in a group of cases of the disease and in appropriate controls. If no association existed between any of the antigens and the disease, the probability that one antigen at least would be associated would be $p = 0.64$.[12] A further example of this analysis might be a healthy person undergoing several health checks and a physician may wish to be sure that the patient was healthy on all of the tests.[12] Second, correction has been suggested in situations where an investigator is searching for significant associations but without a pre-established hypothesis.[12,60] However, this use depends on the 'intention' of the investigator. In an exploratory context, an investigator would not wish to miss a possible effect worthy of further study and therefore, a correction would be inappropriate. However, if the objective was to test everything in the hope that some comparisons would appear significant and the results were not considered to be hypotheses for further study, then a correction should be applied.[3]

## Concluding remarks and advice

Given the problems described and concerns raised by Pernerger[12] what is appropriate statistical advice for authors contemplating multiple statistical testing? As stated by Streiner and Norman[3], to correct or not to correct depends on the circumstances of the study.

1. No correction would be advised in the following circumstances:
   - if the study is restricted to a small number of planned comparisons.[3,61]
   - if a study is exploratory involving *post-hoc* testing of unplanned comparisons which are regarded as hypotheses for further investigation.
   - if multiple usage of a simple test such as '$t$' or '$r$' is envisaged, if it is the results of the *individual* tests that are important. Instead, the exact $p$ values for each individual test should be quoted and discussed appropriately.
   - if it is imperative to avoid a type II error.
2. A Bonferroni correction should be considered if:
   - a single test of the 'universal null hypothesis' ($H_o$) that *all* tests are not significant is required.
   - it is imperative to avoid a type I error.
   - a large number of tests are carried out without preplanned hypotheses in an attempt to establish any results that may be significant.[3]
3. If a correction is required but the original Bonferroni procedure is regarded as too conservative, then a possible alternative is to use the Bonferroni-Holm[62] or Hochberg[63] methods. Both correct for the family-wise

error rate, employ sequential testing, and are less conservative than the original Bonferroni method. In both methods, $p$ values for the various tests are ranked from low to high, i.e., representing the most to the least significant difference. For the Bonferroni-Holm method with comparisons 1 to $T$ and $p = 0.05$ as the '$\alpha$' level, if the most significant $p_1 < \alpha/T$ then the first $H_o$ is rejected. If the first $H_o$ is rejected the analysis is terminated, otherwise the next highest $p$ ($p_2$) is tested. If $p_2 < \alpha/T\text{-}1$ then the second $H_o$ is rejected. The analysis then proceeds to subsequent steps, terminating when a $H_o$ is not rejected.[64] By contrast, the Hochberg method tests the largest $p$ value ($p_T$) first. If $p_T < \alpha$, then all $H_o$ up to and including $p_T$ are rejected and the analysis terminated. If the first $H_o$ is not rejected, the analysis proceeds to test the second highest $p_{T\text{-}1}$ and if $p_{T\text{-}1} < \alpha/2$, all $H_o$ up to and including $p_{T\text{-}1}$ are rejected. The analysis then proceeds to subsequent steps, terminating when a $H_o$ in the sequence is rejected.[64]

4. In all studies involving multiple testing, investigators should clearly describe the design of their study including whether specific hypothesis testing or hypothesis generation is envisaged, provide a rationale for their choice of adjusting or not adjusting $p$ values, justify the method selected if $p$ values are adjusted, and quote adjusted $p$ values correctly.

## Disclosure

The author reports no conflicts of interest and has no proprietary interest in any of the material mentioned in this article.

## References


1. Neyman J & Pearson ES. On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika* 1928; 20A: 175–240.
2. Dunn OJ. Multiple comparison among means. *J Am Stat Assoc* 1961; 56: 52–64.
3. Streiner DL & Norman GR. Correction for multiple testing: Is there a resolution? *Chest* 2011; 140: 16–18.
4. Dmitrienko A & A'Agostino R. Traditional multiplicity adjustment methods in clinical trials. *Stat in Med* ; 32: 5172–5218.
5. Armstrong RA, Davies L, Dunne MCM & Gilmartin B. Statistical guidelines for clinical studies of human vision. *Ophthalmic Physiol Opt* 2011; 31: 123–126.
6. Armstrong RA, Slade SV & Eperjesi F. An introduction to analysis of variance (ANOVA) with special reference to data from clinical experiments in optometry. *Ophthalmic Physiol Opt* 2000; 20: 235–241.
7. Armstrong RA, Eperjesi F & Gilmartin B. The application of analysis of variance (ANOVA) to different experimental designs in optometry. *Ophthalmic Physiol Opt* 2002; 22: 1–9.
8. Armstrong RA & Hilton A. *Statistical Analysis in Microbiology: Statnotes*. Wiley-Blackwell: Hoboken, NJ, 2011.
9. Lam AKC, Lam CH & Chan R. The validity of a digital eyelid tonometer (TGDc-01) and its comparison with Goldmann applanation tonometry-a pilot study. *Ophthalmic Physiol Opt* 2005; 25: 205–210.
10. Morrison KA, Seidel D, Strang NC & Gray LS. The effect of proximity on open-loop accommodation responses measured with pinholes. *Ophthalmic Physiol Opt* 2010; 30: 365–370.
11. Han SJ, Guo Y, Granger-Douetti B, Vicci VR & Alvarez TL. Quantification of heterophoria and phoria adaptation using an automated objective system compared to clinical methods. *Ophthalmic Physiol Opt* 2010; 30: 95–107.
12. Perneger TV. What's wrong with Bonferroni's adjustment. *BMJ* 1998; 316: 1236–1238.
13. Rothman KJ. No adjustments are needed for multiple comparisons. *Epidemiology* 1990; 11: 43–46.
14. Ottenbacher KJ. Quantitative evaluation of multiplicity and public health research. *Am J Epidemiol* 1998; 1477: 615–619.
15. Moyé LA. P-value interpretation and alpha allocation in clinical trials. *Ann Epidemiol* 1998; 86: 351–357.
16. Vision in Preschoolers. (VIP) study group. Impact of confidence number on accuracy of the SureSight vision screener. *Optom Vis Sci* 2010; 87: 96–103.
17. Ehsaei A, Chisholm CM, Pacey IE & Mallen EAH. Off axis partial coherence interferometry in myopes and emmetropes. *Ophthalmic Physiol Opt* 2013; 33: 26–34.
18. Han W, Kuan W, Wang J, Yip SP & Yap M. Influence of eyelid position on wavefront aberrations. *Ophthalmic Physiol Opt* 2007; 27: 66–75.
19. Voges N, Bach M & Kommerell G. Parallactic movement beats binocularity in the presence of external visual noise. *Ophthalmic Physiol Opt* 2012; 32: 308–316.
20. Chou BR, Yuen GS & Dain SJ. Ballistic impact resistance of selected organic ophthalmic lenses. *Clin Exp Optom* 2011; 94: 568–574.
21. Gunvant P, Lievans W, Newman JM III, Gerstier MD, Chang F & Haine CL. Evaluation of some factors affecting the agreement between the proview eye pressure monitor and the Goldmann applanation tonometer measurements. *Clin Exp Optom* 2007; 90: 290–295.
22. Gunvant P, Watkins R, Broadway DC & O'Leary DJ. Repeatability and effects of sequential measurements with POBF tonograph. *Optom Vis Sci* 2004; 81: 794–799.
23. Osubeni EP, Okpala I, Williams TH & Thomas P. Height, weight, body mass index and ocular biometry in patients with sickle cell diseases. *Ophthalmic Physiol Opt* 2009; 29: 189–198.
24. Black AA, Wood JM & Lovie-Kitchen JE. Inferior visual field reductions are associated with poorer functional status

among older adults with glaucoma. *Ophthalmic Physiol Opt* 2011; 31: 282–291.
25. Jinabhai A, Radhakrishnen H, Tromans C & O'Donnell C. Visual performance and optical quality with soft lenses in keratoconus patients. *Ophthalmic Physiol Opt* 2012; 32: 100–116.
26. Lei F, Burns SA, Shao L & Yang Y. Retinal measurements using time domain OCT imaging before and after myopic Lasik. *Ophthalmic Physiol Opt* 2012; 32: 222–227.
27. Cass K & Tromans C. A biometric investigation of ocular components in amblyopia. *Ophthalmic Physiol Opt* 2008; 28: 429–440.
28. Tsujimura S & Tokuda Y. Delayed response of human melanopsin retinal ganglion cells on the pupillary light reflex. *Ophthalmic Physiol Opt* 2011; 31: 469–479.
29. Bueno JM, Cookson CJ, Hunter JJ, Kisilak ML & Campbell MCW. Depolarization properties of the optic nerve head: the effect of age. *Ophthalmic Physiol Opt* 2009; 29: 247–255.
30. Chen D, Lam AKC & Cho P. A pilot study on the corneal biomechanical changes in short-term orthokeratology. *Ophthalmic Physiol Opt* 2009; 29: 464–471.
31. Leat S & Lovie-Kitchen JE. Measuring mobility performance: experience gained in designing a mobility course. *Clin Exp Optom* 2006; 89: 215–228.
32. Debert I, de Alencar LM, Polati M, Souza MB & Alves MR. Oculometric parameters of hyperopia in children with esotropic amblyopia. *Ophthalmic Physiol Opt* 2011; 31: 389–397.
33. Chan B, Cho P & Cheung SW. Repeatability and agreement of two A-scan ultrasonic biometers and ITL Master in non-orthokeratology subjects and post-orthokeratology children. *Clin Exp Optom* 2006; 89: 160–168.
34. Davison P, Aklali M, Loughman J, Scanlon G, Nolan J & Beatty S. Macular pigment: Its association with color discrimination and matching. *Optom Vis Sci* 2011; 88: 816–822.
35. Fenwick E, Xie J, Pesudovs K *et al.* Assessing disutility associated with diabetic retinopathy, diabetes and macular oedema and associated visual impairment using the vision and quality of life index. *Clin Exp Optom* 2012; 95: 362–370.
36. Boost M, Lai S, Ma C & Cho P. Do multipurpose contact lens disinfecting solutions work effectively against non FDA/ISO recommended strains of bacteria and fungi? *Ophthalmic Physiol Opt* 2010; 30: 12–19.
37. Tsai I-L, Tsai C-Y, Kuo L-L, Liou S-W, Lin S & Wang I-J. Transient changes of intraocular pressure and anterior segment configuration after diagnostic mydriasis with 1% tropicamide in children. *Clin Exp Optom* 2012; 95: 166–172.
38. Cho P, Cheng SY, Chan NY & Yip WK. Soft contact lens cleaning: rub or no rub. *Ophthalmic Physiol Opt* 2009; 29: 49–57.
39. Flores-Rodriguez P, Gili P & Martin-Rios MD. Sensitivity and specificity of time-domain and spectral domain optical coherence tomography in differentiating optic nerve head drusen and optic disc oedema. *Ophthalmic Physiol Opt* 2012; 32: 213–221.
40. Kemper A, Bruckman D & Freed G. Prevalence and distribution of corrective lenses among school-age children. *Optom Vis Sci* 2004; 81: 7–10.
41. Myint J, Edgar DF, Kotecha A, Murdoch IE & Lawrenson JG. Barrier perceived by UK-based community optometrists to the detection of primary open angle glaucoma. *Ophthalmic Physiol Opt* 2010; 30: 847–853.
42. Campbell JL, Griffin L, Spalding AB & Mir FA. The effect of abnormal colour vision on the ability to identify and outline coloured clinical signs and to count stained bacilli in sputum. *Clin Exp Optom* 2005; 88: 376–381.
43. Reiter C, Leisang D & Madson E. Survey of German clinical prescribing philosophies for hyperopia. *Optom Vis Sci* 2007; 84: 131–136.
44. Hsu S-Y, Ko M-L, Linn G, Chang M-S, Shen M-M & Tsai RK. Effects of age and disc area on optical coherence tomography measurements and analysis of correlations between optic nerve head and retinal nerve fibre layer. *Clin Exp Optom* 2012; 95: 427–431.
45. Court H, Greenland K & Margrain TH. Predicting state anxiety in Optometric practice. *Optom Vis Sci* 2009; 86: 1295–1302.
46. Mash C & Dobson V. Intraobserver reliability of the Teller acuity card procedure in infants with perinatal complications. *Optom Vis Sci* 2005; 82: 817–822.
47. Lee TT & Cho P. Relative peripheral refraction in children: twelve-month changes in eyes with different ametropias. *Ophthalmic Physiol Opt* 2013; 33: 283–293.
48. Lamoureux E, Ecosse L, Ferraro JG *et al.* Are standard instruments valid for the assessment of quality of life and symptoms in glaucoma. *Optom Vis Sci* 2007; 84: 789–796.
49. Oliveira-Soto L & Efron N. Morphology of corneal nerves in soft contact lens wear. A comparative study using confocal microscopy. *Ophthalmic Physiol Opt* 2003; 23: 163–174.
50. Cufflin MP, Hazel CA & Mallen EAH. Static accommodative response following adaptation to differential levels of blur. *Ophthalmic Physiol Opt* 2007; 27: 353–360.
51. Yang Y & Wu F. Comparison of the wavefront aberrations between natural and pharmacological pupil dilation. *Ophthalmic Physiol Opt* 2007; 27: 220–223.
52. Nickla D, Zhu X & Wallman J. Effects of muscarinic agents in chick choroids in infarct eyes and eye cusps: evidence for a muscarinic mechanism in choroidal thinning. *Ophthalmic Physiol Opt* 2013; 33: 245–256.
53. Yu M, Keintz MA, Thomas ML, Johnson D, Hotchkiss ER & Rosso MB. Operational implications of varying ambient light levels and time-of-day effects on saccadic velocity and pupillary light reflex. *Ophthalmic Physiol Opt* 2007; 27: 130–141.
54. Harle DE & Evans BJW. Subtile binocular vision anomalies in migraine. *Ophthalmic Physiol Opt* 2006; 26: 587–596.
55. Januscheit S, Doughty MJ & Button NF. On the use of Orbscan II to assess the peripheral corneal thickness in humans: a comparison with ultrasound pachymetry measures. *Ophthalmic Physiol Opt* 2007; 27: 179–189.

56. Chen D, Lam AKC & Cho P. Posterior corneal curvature change and recovery after 6 months of overnight orthokeratology treatment. *Ophthalmic Physiol Opt* 2010; 30: 274–280.
57. Lee S-Y, Petznick A & Tong L. Associations of systemic diseases, smoking and contact lens wear with severity of dry eye. *Ophthalmic Physiol Opt* 2012; 32: 518–526.
58. Richdale K, Mitchell GL & Zadnik K. Comparison of multifocal and monovision soft contact lens corrections in patients with low-astigmatic presbyopia. *Optom Vis Sci* 2006; 83: 266–273.
59. Aakre BM, Ystenaes AE, Doughty MJ, Austrheim O, Westerfjell B & Lie MT. A 6-month follow up of successful refits from daily disposable soft contact lenses to continuous wear of high-Dk silicone-hydrogel lenses. *Ophthalmic Physiol Opt* 2004; 24: 130–141.
60. Golebioswki B, Lim M, Papas E & Stapleton F. Understanding the stimulus of an air-jet aesthesiometer: comparative modelling and subjective interpretation. *Ophthalmic Physiol Opt* 2013; 33: 104–113.
61. Schulz KF & Grimes DA. Multiplicity in randomised trials 1: endpoints and treatments. *Lancet* 2005; 3659470: 1591–1595.
62. Holm S. A simple sequential rejective multiple test procedure. *Scand J Statist* 1979; 62: 65–70.
63. Hochberg Y & Tamhane AC. *Multiple Comparison Procedures*. Wiley: New York, NY, 1987.
64. Huang Y & Hsu JC. Hochberg's step-up method: cutting corners off Holm's step-down method. *Biometrika* 2007; 94: 965–975.

**Richard A. Armstrong** was educated at King's College London and St. Catherine's College Oxford. He has been a lecturer in Botany, Microbiology, Ecology, Neuroscience, and Optometry during his 36 years at Aston. His research interests include the neuropathology of neurodegenerative diseases with special reference to vision and the visual system. He also has a particular interest in the application of statistical methods in research.