Graefe's Arch Clin Exp Ophthalmol
DOI 10.1007/s00417-006-0474-4

CLINICAL INVESTIGATION

The Freiburg Visual Acuity Test-Variability unchanged by post-hoc re-analysis

Michael Bach

Received: 26 January 2006 / Revised: 28 August 2006 / Accepted: 9 October 2006

Abstract
Background The Freiburg Visual Acuity and Contrast Test (FrACT) has been further developed; it is now available for Macintosh and Windows free of charge at http://www.michaelbach.de/fract.html. The present study sought to reduce the test-retest variability of visual acuity on short runs (18 trials) by post-hoc re-analysis.
Methods The FrACT employs advanced computer graphics to present Landolt Cs over the full range of visual acuity. The sequence of optotypes presented follows an adaptive staircase procedure, the Best-PEST algorithm. The Best-PEST threshold obtained after 18 trials was compared to the result of a post-hoc re-analysis of the acquired data, where both threshold and slope of the psychometric function were estimated via a maximum-likelihood fit.
Results Testing time was 1.7 min per run on average. Test-retest reproducibility was ±2 lines (or ±0.2 logMAR) for a 95% confidence band (using 18 optotype presentations per test run). Post-hoc psychometric fitting reproduced the Best-PEST result within 1%, although the individual slopes varied widely; test-retest reproducibility was not improved.
Conclusions The FrACT offers advantages over traditional chart testing with respect to objectivity and reliability. The similarity between the results of the Best-PEST vs. post-hoc analysis, fitting both slope and threshold, suggest that there is no disadvantage to the constant slope assumed by Best PEST. Furthermore, since variability was not reduced by post-hoc analysis, for high reliability more trials should be employed than the 18 trials per run used here.

Keywords Visual acuity · Automation · Computer

M. Bach (✉)
University Augenklinik Freiburg,
Killianstr. 6,
79106 Freiburg, Germany
e-mail: michael.bach@uni-freiburg.de

Introduction

It is surprising that the signal detection theory [9] and computer assistance have not been widely exploited for the measurement of visual acuity (VA), although these methods provide a considerable reduction of confounding influences. In contrast, clinical perimetry has extensively utilized these methods. In an attempt to fill this gap, the Freiburg Visual Acuity and Contrast Test (FrACT) was developed [1]. The use of computer graphics and interactive computer assistance for the measurement of VA appeared quite straightforward. However, unexpected hurdles appeared, including (1) sufficient resolution of display devices, (2) good automatic bracketing of the threshold, (3) problems of automatically dealing with the vagaries of patient responses, and (4) unrealistic expectations of the stability of acuity measures, even if obtained quasi-objectively. Since then, FrACT has been used in a number of studies, while being continuously further developed based on the feedback of numerous vision researchers, optometrists and ophthalmologists. Other approaches emulated the ETDRS charts [18] or were variants of the FrACT approach [16]. ETDRS and FrACT results were recently compared in this laboratory and found to agree closely [19]. The present article will describe the current version of the FrACT and a test of the viability of the constant slope assumption of the Best-PEST threshold estimation paradigm. It was hoped that post-hoc processing could reduce the test-retest variability, so that briefer test runs could be employed.

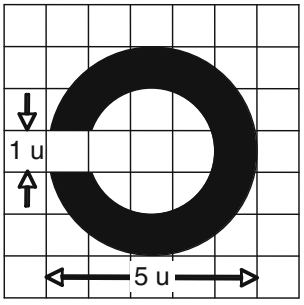

Fig. 1 Landolt C. The unit *u*, measured in minutes of arc, defines the decimal visual acuity (*VA*): For a VA of 1.0 (=20/20), *u* would span 1 min of visual angle

Methods

Versions

FrACT is currently available in two different versions: (1) the old version, which runs only on Macintosh computers, but also offers contrast testing, and (2) the universal version, the latter being programmed such that executables for two major computer platforms are produced (Macintosh and Windows) and having a web-browser plug-in that also runs on Linux. The two versions are identical with respect to the optotype geometry, calibration and threshold estimation algorithm; they differ in the programming environment and optional capabilities. The present report concentrates on acuity assessment with the second, more recent, platform-agnostic version.

Equipment

Nearly any kind of computer manufactured in the last 5 years can be employed. The computational requirements are very low relative to current computer capabilities. The graphics should be able to resolve at least 256 gray levels, or millions of colors (3×8-bit color depth). The resolution of the visual display unit (VDU) is the most likely bottleneck (see "Limitations" in Discussion). Both CRT- or LCD-type VDUs are possible. EN ISO 8596 [6] details a luminance of the Landolt C between 80 and 320 cd/m^2 at a contrast of 85%. This is easily reached with consumer-grade equipment. The present results were obtained with a 17" CRT monitor at a distance of 4 m, a luminance of 105 cd/m^2 at a contrast of 95% and a background illuminance of 60 lux, measured at the subject's eye.

Nomenclature and optotypes

Visual acuity will be abbreviated with VA and defined operationally; the decimal notation will be used throughout [FrACT results are also optionally displayed as Snellen ratio or logMAR = −log(VA)]. The operational definition of VA is as follows: VA=1/d_t where d_t= the threshold gap size (minutes of arc). The definition of threshold will be given in the next section. The term gap refers to the gap of the Landolt C (Fig. 1). Landolt C optotypes are presented on the computer screen. The relation of the outer and inner diameter to the gap size is shown in Fig. 1. The independent variable is the gap size. Resolution is improved by using anti-aliasing [2, 8], allowing sub-pixel rendering of the Landolt C, and for non-integer gap sizes (e.g., 1.25 pixels). The gap can be presented in one of eight

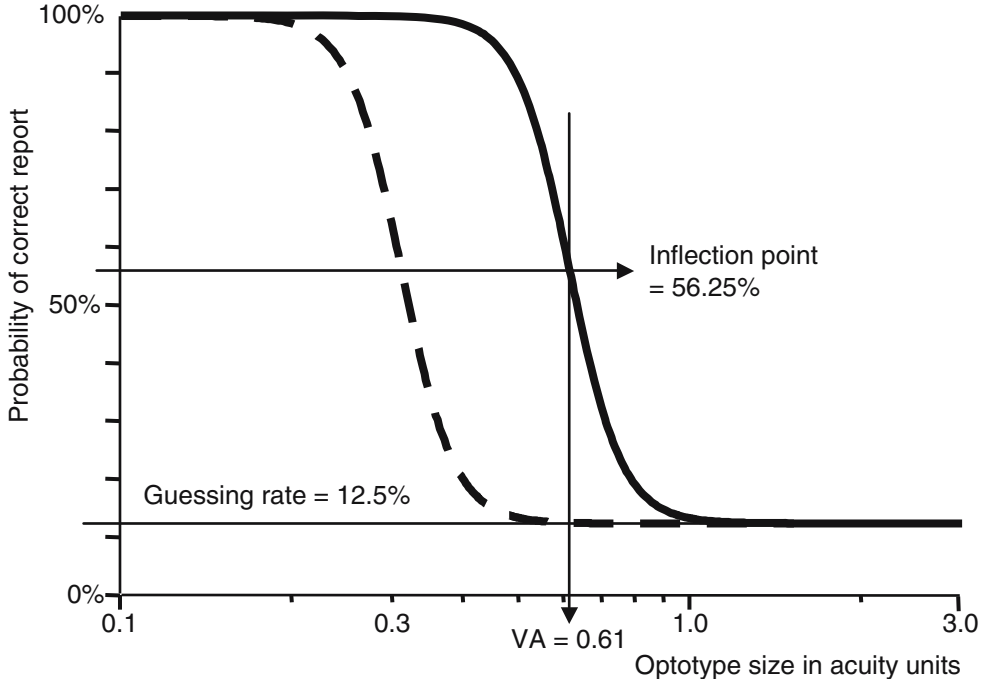

Fig. 2 The psychometric function governing acuity measurements, here shown for a subject with decimal VA=0.61. When the optotypes are large (*left*), they will be recognized correctly. With decreasing optotype size, the probability to report the gap direction correctly will decrease, eventually down to the guessing rate, which equals 12.5% for eight possible gap directions. In between, the detection rate is well described by a logistic function. The acuity is defined at the inflection point, the middle between 100% correct and the guessing rate. For subjects with differing acuity, this psychometric function shifts horizontally; the *dashed curve* example corresponds to VA=0.3

or in one of four positions; in the present study, eight different orientations were used.

Threshold definition

Like all sensory thresholds, the detection rate versus optotype size is described by a psychometric function (Fig. 2) [11]. Given this continuous relation, threshold definition is not obvious. The optimal choice from a signal-detection point of view is the point of steepest slope, which is also the point of inflection. With the use of eight alternatives, this point lies in the middle between the guessing rate of 12.5% and 100%, i.e., at 56.25%. At this point, any deviation on the detection scale (ordinate) transforms into the smallest possible deviation on the acuity scale (abscissa). This definition is widely used and also underlies the EN ISO 8596 standard [6]. One could also define this region as the most uncomfortable one for the patient: here, they are most uncertain whether or not they can recognize the target.

Threshold estimation

It is suggested by signal detection theory [9], and has been shown experimentally (e.g., [17]), that when comparing the psychometric acuity function in subjects with a wide range of acuity, the position of the inflection point shifts, but slope stays rather constant, when plotted on a log(VA) scale. Neglecting lapses, only one parameter needs to be estimated, namely the threshold. A number of algorithms have been developed for such a situation (review: [20] or the special edition of *Perception* [15]). For FrACT, the Best-PEST algorithm was selected [12], which needs no prior information and assumes a constant fixed slope of the psychometric function. In its purest form, this algorithm always presents optotype sizes at the currently most likely threshold, because that maximizes information gain. This most likely position is calculated by a maximum likelihood procedure that looks for the position of the psychometric function that would reproduce all previous trial results with the highest likelihood.

Practically, a number of modifications are useful. The initial optotype size depends on the highest and lowest possible acuity (the full range); thus, it would be strongly influenced by screen size, pixel size and observation distance. To avoid such influences, the first four trials present optotypes corresponding to an acuity of 0.1, 0.2, 0.4 and 0.8 (following EN ISO 8596 [6]) as long as they are all correctly discriminated. From then on, the Best PEST takes over. As a boundary condition, the optotype never becomes smaller than 5 pixels in diameter (corresponding to a gap size of 1 pixel), as—even with anti-aliasing—the effective contrast would vanish below this limit. Any non-integer value of the gap size ≥1 pixel is possible though, with the help of anti-aliasing. The Landolt C orientation is calculated randomly for each trial.

An important parameter is the number of trials [13]. In the interest of rapid measurement and avoidance of subject fatigue, it should be as low as possible; for best precision, however, it should be as high as possible. For clinical studies, 30 trials yield a test-retest comparable to the ETDRS procedure [19]. In the present study, only 18 trials were presented. This increases variability, making room for improvement. Post-hoc re-analysis (see below) sought to reduce the variability, though without success.

Procedure

The procedure with FrACT is subject to the same boundary conditions as any acuity test: well-defined surround luminance, no screen illumination that would reduce contrast, a

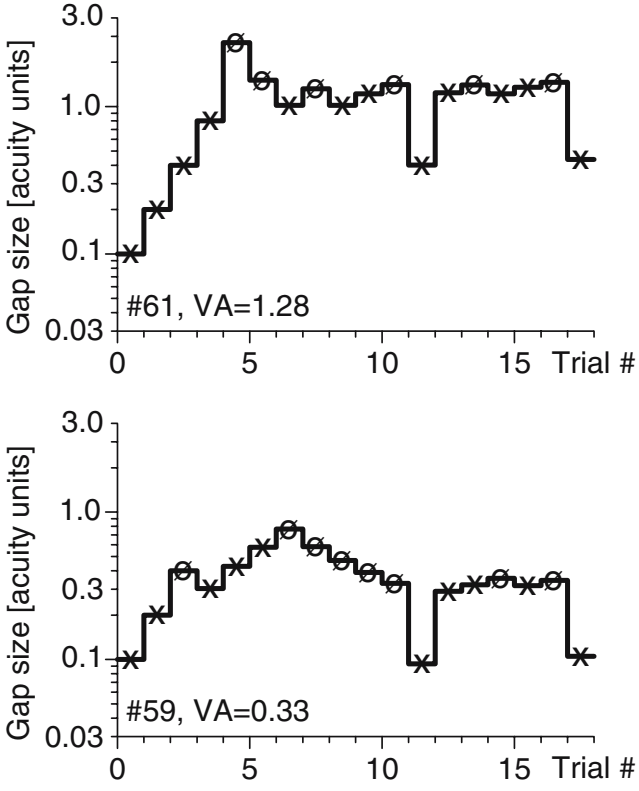

Fig. 3 Two representative runs of FrACT in two subjects. The ordinate represents the Landolt-C gap size; the abscissa covers the sequence of 18 trials. Trials with correct responses are indicated by ×, incorrect ones by ∅. The low-acuity outliers at trial 12 and 18 represent the bonus trials. In the top subject, the initial 0.1–0.8 acuity sequence was recognized correctly, but the presentation at nearly 3.0 VA was not correctly identified; the algorithm subsequently homed in on VA=1.28. In the bottom subject, the Landolt C corresponding to 0.4 acuity was not correctly identified with a subsequent step-down in acuity, with: ν, acuity value of the optotype; p_{chance}, guessing probability =0.125

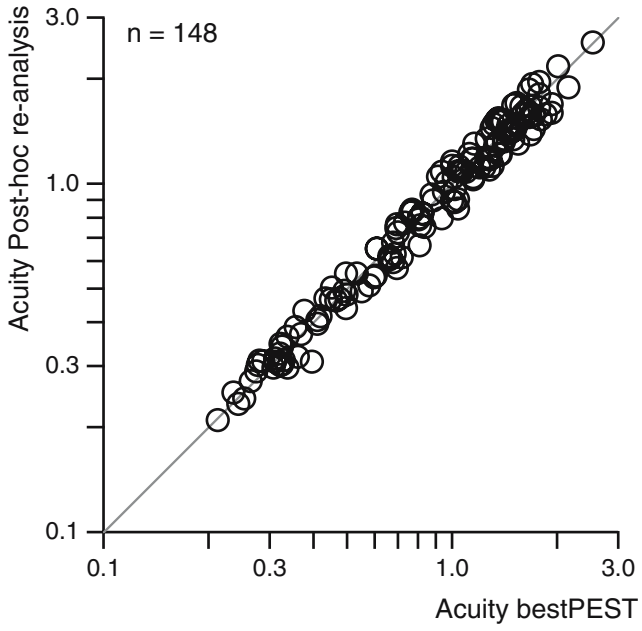

Fig. 4 Comparison of acuity results obtained by Best-PEST and post-hoc fitting of a psychometric function with slope and threshold as free parameters

calibrated setup, defined observer distance, appropriate correction and adequate training of examiner and subject.

To obtain data for the present study, FrACT was run in 74 eyes of 37 subjects (age range, 19–71 years, mean 38 years) without known eye disease, using eight different optotype orientations. Most performed the test for the first time. Their correction was not necessarily the best, which is not a problem here, since test-retest variability was the target variable. The subjects operated a remote response box, with keys labeled by Landolt Cs with appropriate orientations. The test was explained, one binocular run performed to familiarize the subject with the procedure (these results were discarded), then four runs in the sequence left eye/right/right/left were recorded; for every alternate subject, the sequence of eyes was inverted. The data of every run were saved using the "export to clipboard" option (see manual [3]). To extend the range of acuities to lower values, some subjects were assessed in a second session, about 2 weeks later, with glasses degrading optical quality. The optical quality was not reduced with plus lenses because it is likely that accommodation will be unstable under such a condition. The commercially available Bangerter transparencies proved too unequal across their surface. Prompted by a personal communication by Bernhard Rassow, I experimented with plastic transparencies manufactured as office paper binders until one was found with an even distribution of its 'milky' quality and the right amount of degradation, and mounted cutouts in a cardboard trial frame. Such a scatter transparency, due to its high amount of wide-angle scatter, models some media opacities [10].

Most subjects can perform the test in a self-paced mode, especially on repeat visits. It is useful to help with occasional supportive statements indicating that "errors" are entirely in order, and not to ponder the response too long. This self-paced mode can allow the examiner to pursue other tasks in the mean time.

Post-hoc processing

Since inception of this program, computer speed has risen by several orders of magnitude. Thus, the psychometric paradigm needs no longer be chosen for speed, which had been a major design goal for the Best-PEST algorithm [12], and can be re-evaluated. Best PEST assumes a constant slope and zero lapse rate.

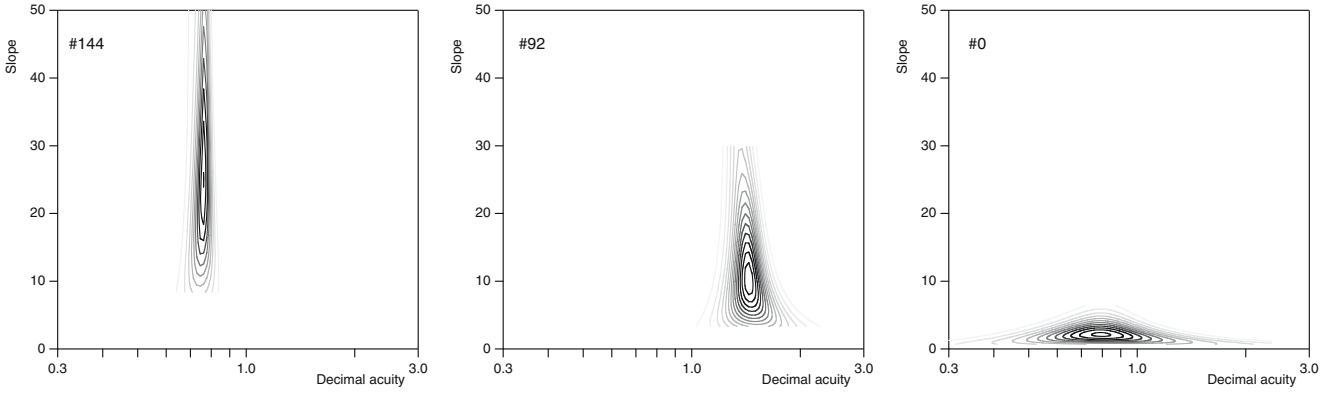

Fig. 5 Typical characteristics of the psychometrical function (cf. Fig. 2) for three different subjects. The z-axis (represented via *contour lines*) depicts the likelihood of the fit producing the specific test run result; the z-ranges covered are: $0–5{\cdot}10^{-3}$, $0–3{\cdot}10^{-3}$, $0–6{\cdot}10^{-5}$, respectively. The abscissa represents the acuity threshold, and slope is on the ordinate.The two *left graphs* are typical for low and high acuity cases; the right one depicts a run where the subject gave an incorrect response to a (very easy) bonus trial, resulting in a low slope. In none of the 148 cases was there a marked obliqueness of the likelihood 'hill' (the *center graph* represents an extreme case), indicating that the threshold estimate does not depend strongly on the slope

The run data thus obtained (similar to those depicted in Fig. 3) were fitted with a psychometric function P using maximum-likelihood [21, 24] where both the threshold v_0 and slope s were free parameters according to the following Eq. (1):

$$P(\nu)=p_{chance}+(1-p_{chance})/(1+(\nu_0/\nu)^s)) \quad (1)$$

To asses the quality of these post-hoc acuity values, their test-retest variability was calculated. To quantify test-retest variability, often the correlation coefficient is calculated. This, however, depends strongly on the total range, and one can achieve high values just by including endpoint values. The mean coefficient of variation (CV) across every test-retest pair does not suffer from this limitation and was chosen instead. There is one problem using the CV though: VA is not normally distributed, but logVA or logMAR are. But the CV calculation cannot be applied to data crossing zero, which is the case with both logVA and logMAR. Since the test-retest values are rather close (around 10%), and, as mathematicians say, everything is linear to the first order, the CV was calculated on the VA scale.

Results

In Fig. 3, two representative runs of FrACT in two subjects are depicted. It can be seen that a run starts with "easy" optotypes (low acuity) that become smaller until an incorrect response is encountered (the fifth trial for the upper example). Consequently, the Best-PEST algorithm next selects a somewhat larger optotype.

The average time per 18-trial run, including data transfer to a spreadsheet, was 103±40 s (range, 53–210 s). It took about 6 min for the acuity of both eyes including a binocular training run.

Figure 4 demonstrates that the threshold obtained by fitting the psychometric function with slope as another free parameter never differed by more than one line (1 dB, or a factor of 1.26) from the one obtained by Best PEST. On average, the results differed by 1.1%. The slope ranged widely from 2.0 to over 100, averaging at 17.0; this is markedly higher than the slope value of 1.7 currently used. Figure 5 depicts the maximal likelihood function depending on slope and threshold (decimal acuity) for three representative cases. On the left, the slope of the psychometric function is very steep, while on the right it is very shallow (in that run, an incorrect response to a large optotype, a bonus trial, was entered). The threshold depends very little on slope as can be seen from the missing obliqueness of the likelihood vs. slope and threshold function. A brief explanation of the likelihood function may be in order: Likelihood is the hypothetical probability that an event that has already occurred will yield a specific outcome. The concept differs from that of a probability in that a probability refers to the occurrence of future events, while a likelihood refers to past events with known outcomes

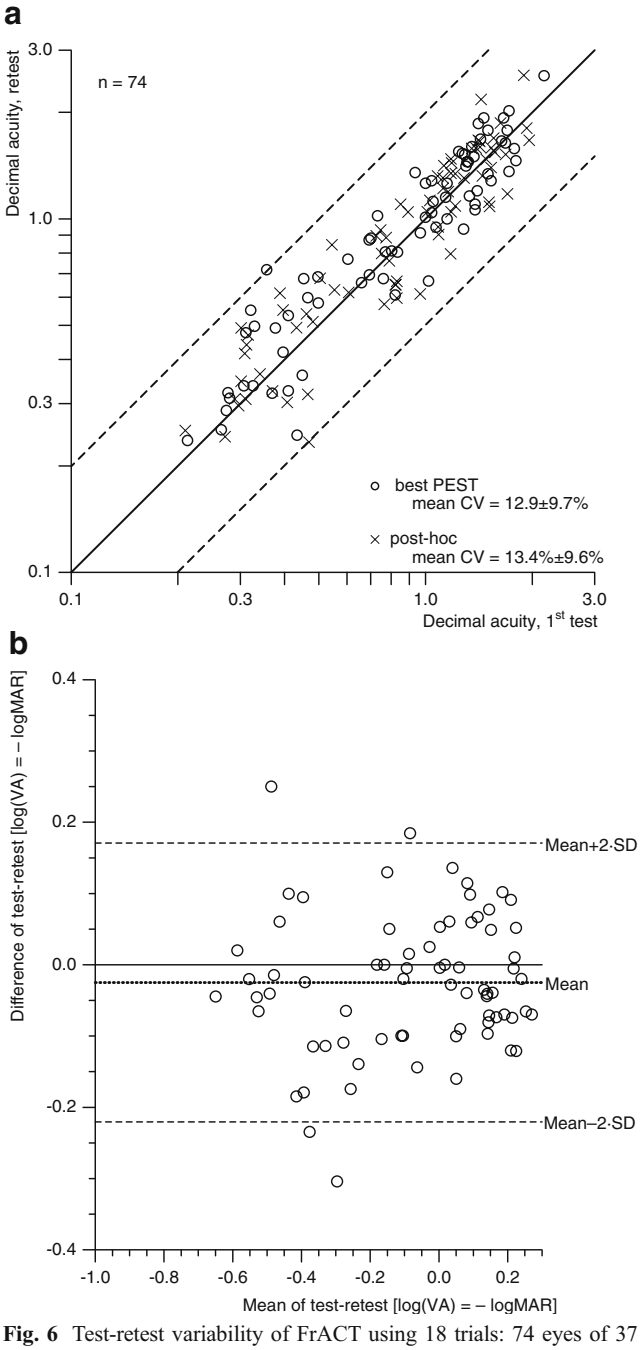

Fig. 6 Test-retest variability of FrACT using 18 trials: 74 eyes of 37 naive, visually normal subjects, not necessarily wearing best correction, were analyzed. (**a**) Scatter plot. Along the continuous 45° line, perfect reproducibility would be obtained. The *parallel dashed lines* indicate a deviation of ±3 lines on retest from the initial test. There was no improvement of repeatability by post-hoc analysis. (**b**) The best-PEST data, depicted as difference test-retest vs. average of test-retest (Bland-Altman plot [4, 5]); the abscissa covers the same range as (**a**). The mean test-retest difference is close to zero, and variability does not change markedly across the acuity range covered

[22]. The event here is the occurrence of the entire sequence of correct-incorrect responses, given the specific values of acuity threshold and slope value (cf. Fig. 1).

Figure 6 illustrates the test-retest reproducibility. Points on the continous 45° line would be perfectly reproduced; the dashed lines correspond to deviations by a factor of two (corresponding to three lines on an acuity chart or 3 dB). For the Best-PEST method, 72 of 74 (97.3%) run pairs differed by 2 dB or less; the mean CV was 12.9±9.7%. After post-hoc processing, 73 of 74 (98.6%) run pairs differed by 2 dB or less; the mean CV was 13.5±9.7%. Thus, post-processing does not yield significantly different thresholds (P=0.79, Wilcoxon test), it does not reduce test-retest variability, and it removes about as many outliers as it adds. In Fig. 6b, the Best-Pest results are depicted as a Bland-Altman [4, 5] plot (after taking the logarithm of all acuities, the difference test-retest vs. average of test-retest). The mean difference (dotted line) was 0.025 logMAR, and the dashed lines indicate ±2·SD around the mean (2·SD= 0.196 logMAR). This plot shows that: (1) The negative mean difference hints at a small, though insignificant (P= 0.6) learning effect, and (2) there is no marked skewness for low acuities, but a tendency towards higher variability. The post-hoc data have been left out to avoid clutter; the mean difference was 0.014 logMAR and the 95% confidence band is spanned by 0.204 logMAR.

Discussion

The old version of FrACT has been successfully validated in independent laboratories [7, 14, 23]. The new version is geometrically identical and showed an agreement between the (new) FrACT and ETDRS charts within 9% down to very low acuities in the author's laboratory [19]. This suggests that the FrACT results are bias-free estimators of visual acuity over the full range from ≈0.01 to ≈3.0. The present study assessed the test-retest variability of FrACT with only 18 trials. The test-retest variability as quantified by the coefficient of variation (CV) of VA was around 13%, corresponding to about half a line (1 line = a factor of 1.26). The corresponding 95% confidence interval spans ±0.196 logMAR. This leaves room for improvement.

The post-hoc analysis, namely fitting slope as another free parameter in addition to the threshold, resulted in nearly identical acuity estimates and nearly identical average test-retest values. The maximum likelihood fitting surface showed that slope and threshold are highly decoupled; in other words, whichever value of slope is chosen has very little influence on the acuity outcome. This suggests that the fixed slope as used in Best PEST is no disadvantage. Somewhat disappointingly, post-hoc processing did not improve test-retest variability. Either this reflects inherent fluctuation of the threshold itself, or the loss of degrees of freedom to estimate the additional parameter slope offsets a possible closer approximation of the psychometric function. Still, post-hoc processing has been integrated into FrACT as an option.

A number of additional modifications of the post-hoc analysis were tried out: removal of the bonus trial results, restricting analysis to the final part, iteratively removing outliers, and removing erroneous bonus trials. None of these modifications resulted in a lower test-retest variability.

One problem of FrACT occurs when a subject mistypes their first response; that is when a very large optotype is seemingly not recognized correctly. The Best-PEST algorithm then searches too long for the threshold in the low acuity region and may not converge to full acuity. In such a case, it is best to abort the run and restart. This was not necessary in the present study.

The main application fields for FrACT are thus clinical studies where acuity is an outcome variable. FrACT can be seen as an automated alternative to ETDRS, extending its range both at the upper and lower end and being safe from being learned by heart on repeated testing. In laboratory environments, FrACT has proven useful for subject screening and for quantifying acuity after optical or physiological manipulations. Since the present study was not successful in reducing the variability of the rather short 18-trial runs, for highly reliable results, the test should either be repeated and the results averaged, or the number of trials should be increased to 30 [1, 19].

Acknowledgement Thanks to many users for their inspiring support, providing feedback that helped to root out bugs and suggesting useful expansions. Special thanks to Lew Harvey, Hans Strasburger and Thomas Meigen for tutoring in signal detection theory, psychometric threshold assessment and probability statistics and to Margret Schumacher for assiduous testing. Finally, thanks to two very persistent reviewers who considerably helped to clarify my thoughts.

References

1. Bach M (1996) The Freiburg Visual Acuity Test-automatic measurement of visual acuity. Optom Vis Sci 73:49–53
2. Bach M (1997) Anti-aliasing and dithering in the Freiburg Visual Acuity Test. Spat Vis 11:85–89
3. Bach M (2006) Homepage of the Freiburg Visual Acuity and Contrast Test ('FrACT'). Retrieved 2006-07-04, from http://www.michaelbach.de/fract.html
4. Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1:307–310
5. Bland JM, Altman DG (1995) Comparing methods of measurement: why plotting difference against standard method is misleading. Lancet 346:1085–1087
6. CEN (Comité Européen de Normalisation) (1996) Ophthalmic optics-visual acuity testing-the standard optotype and its presentation. Beuth-Verlag, Berlin

7. Dennis RJ, Beer JM, Baldwin JB, Ivan DJ, Lorusso FJ, Thompson WT (2004) Using the Freiburg Acuity and Contrast Test to measure visual performance in USAF personnel after PRK. Optom Vis Sci 81:516–524
8. Foley JD, Van Dam A, Feiner SK, Hughes JF (1990) Computer Graphics, Principles and Practice. Addison-Wesley, Reading
9. Green DM, Swets JA (1966) Signal detection theory and psychophysics. Wiley, New York
10. Hess R, Woo G (1978) Vision through cataracts. Invest Ophthalmol Vis Sci 17:428–435
11. Klein SA (2001) Measuring, estimating, and understanding the psychometric function: a commentary. Percept Psychophys 63:1421–1455
12. Lieberman HR, Pentland AP (1982) Microcomputer-based estimation of psychophysical thresholds: The best PEST. Behav Res Methods Instrument 14:21–25
13. Linschoten MR, Harvey LO, Jr., Eller PM, Jafek BW (2001) Fast and accurate measurement of taste and smell thresholds using a maximum-likelihood adaptive staircase procedure. Percept Psychophys 63:1330–1347
14. Loumann Knudsen L (2003) Visual acuity testing in diabetic subjects: the decimal progression chart versus the Freiburg visual acuity test. Graefes Arch Clin Exp Ophthalmol 241:615–618
15. Macmillan NA (2001) Threshold estimation: the state of the art. Percept Psychophys 63:1277–1278
16. Peters BT, Bloomberg JJ (2005) Dynamic visual acuity using "far" and "near" targets. Acta Otolaryngol 125:353–357
17. Petersen J (1990) Zur Fehlerbreite der subjektiven Visusmessung. Fortschr Ophthalmol 87:604–608
18. Ruamviboonsuk P, Tiensuwan M, Kunawut C, Masayaanon P (2003) Repeatability of an automated Landolt C test, compared with the early treatment of diabetic retinopathy study (ETDRS) chart testing. Am J Ophthalmol 136:662–669
19. Schulze-Bonsel K, Feltgen N, Burau H, Hansen LL, Bach M (2006) Visual acuities "Hand Motion" and "Counting Fingers" can be quantified using the Freiburg Visual Acuity Test. Invest Ophthalmol Vis Sci [in print]
20. Treutwein B (1995) Adaptive psychophysical procedures. Vision Res 35:2503–2522
21. Treutwein B, Strasburger H (1999) Fitting the psychometric function. Percept Psychophys 61:87–106
22. Weisstein EW. (2006) "Likelihood." From MathWorld - A Wolfram Web Resource. Retrieved 2006-06-27, from < http://mathworld.wolfram.com/Likelihood.html >
23. Wesemann W (2002) [Visual acuity measured via the Freiburg visual acuity test (FVT), Bailey Lovie chart and Landolt Ring chart]. Klin Monatsbl Augenheilkd 219:660–667
24. Wichmann FA, Hill NJ (2001) The psychometric function: I. Fitting, sampling, and goodness of fit. Percept Psychophys 63:1293–1313