EBU Technical Review : No. 282 (March 2000)

Picture quality in MPEG video

J. Fletcher and M. Prior-Jones
BBC Research & Development

This article describes an investigation made into the effects of DCT coefficient quantization on the picture quality of MPEG-coded video. This involved subjective tests in which viewers were asked to grade pictures that had been coded at a fixed level of quantization. The results give a relationship between subjective picture quality and *quantizer_scale*.

Introduction

MPEG-2 encoding has been widely adopted as a digital video compression system, and is the basis of the new “digital” television services. It is, however, still a relatively new technology and there are still questions to be answered about how to use it effectively. One of the chief complaints about MPEG-2-compressed video is the difficulty of assigning bit-rates to programmes. High bit-rates give better quality, but reduce the number of programmes per multiplex. Picture quality in an MPEG-2-based system is dependent on both the available bit-rate and the content of the programme being shown. Programmes with fast-moving action and frequent shot-changes look worse than “talking heads” for the same bit-rate, because of the nature of the compression system. It was thought that a programme-independent measure of picture quality would be useful, as it could be used as a guide when allocating bit-rates, or as a determining factor in statistical multiplexing algorithms.

Although the MPEG-2 system uses a large number of techniques to compress video, only two are actually lossy: the 4:2:2 to 4:2:0 chrominance subsampling and the quantization of the DCT coefficients. The latter can be varied from very fine (where the artefacts are barely perceptible) to very coarse (where the artefacts are very annoying), and they directly affect the output bit-rate (coarser = lower bit-rate). As MPEG-2 is usually used in a constrained bit-rate environment, a feedback loop is built into the encoder. The encoder fills a buffer at a variable rate, which is emptied by the transmission channel at a constant rate. If the buffer begins to fill up, the quantization is made coarser. If the buffer empties, the quantization is made finer. In this way a constant bit-rate transmission is produced. The same feedback loop operates in a statistical multiplexing system but the transmission channel bit-rate (for this programme) is varied from time to time according to the demands of other programmes in the multiplex.

The idea behind this series of experiments was to determine the relationship between *quantizer_scale* (the variable which controls the coarseness of DCT coefficient quantization) and the actual picture quality. We hoped to find a threshold point, beyond which most people could not distinguish between encoded and uncoded pictures, as this could be used to provide an "opportunistic data service". This technique involves setting a lower limit on the quantizer, so that the programme is never coded at a level finer than the viewer can see. This reduces the bit-rate and frees up data capacity that could be used for a non-time-critical data service (such as teletext).

The equipment used

The coder available to us was the one developed by the COUGAR [1] project, which could readily be customised to fix the quantizer and vary the bit-rate, as opposed to vice-versa. This coder is somewhat unusual (it is a research prototype rather than a commercial product), and uses the BBC/Snell & Wilcox Phase Correlation Motion Estimation technique, rather than the block-matching used by other coders [2].

The coder was set up with a minimum-slice quantizer setting and a maximum output bit-rate (15 Mbit/s, the maximum possible). This meant that the coder fixed the slice quantization at the minimum setting unless the output bit-rate approached 15 Mbit/s, where it coarsened the quantization to avoid a buffer overflow. We did experience this effect with certain programme items which, when coded at *quantizer_scale* = 6, required more than 15 Mbit/s (the maximum the coder could deliver) and thus had short bursts of coding at a coarser quantization. Most of these sequences were not used in the final tests; the only one in the test to be affected was the sequence showing a rugby match.

The coder's "adaptive quantization" feature (which adjusts the quantization of individual macroblocks) was switched on. This coarsens the quantization on fast-moving areas and sharpens it on plain areas and on areas with a high red content. This helps to even out the picture quality between the programmes.

[1] COUGAR was a project in the EU RACE programme.

[2] Phase correlating coders code the signal using the actual motion vectors of the video, whereas block-matching coders search for the best fit within a limited search range. These techniques are alternatives, and each has its own strengths and weaknesses.

Preliminary experiments

Several experiments were performed with the coder to determine the parameters and types of programme to use. Initially, the coder was set up in the lab with an off-air feed of BBC 1 and various settings were tried with various programmes. Programmes that gave interesting results were acquired on tape from the Archive. These programmes (being "daytime TV") were sourced in PAL format on a D3 tape. We wanted some component-recorded programmes as well, and a number were selected from the BBC's weekly programme guide, the *Radio Times*, (in which they're identified as *Digital Wide-*

Table 1
The programme items used

Title	Format	Description
Box Hill	Originally 1250-line widescreen HD, but converted to 625-line 4x3.	Helicopter flight over wooded hillside.
Eastenders	PAL 4x3	Popular soap opera. Clip showing woman, indoors, talking to her husband and folding a sequinned dress.
GW Garden	Originally 16x9 component, but converted to 4x3	An extract from *Gardener's World*, showing Stephen Lacey walking through a garden.
GW Potting Up	Originally 16x9 component, but converted to 4x3	Alan Titchmarsh potting up outside his greenhouse.
Heart By-pass	Originally widescreen film, but recorded as 4x3 PAL	Jonathan Meades documentary about Birmingham. Clip shows him taking toy cars out of a washing machine in a laundrette.
Ironside	Originally film, but recorded as 4x3 component.	1970's American cop show. Clip shows a boy walking down a crowded street.
Playdays	PAL 4x3	Children's programme. Clip shows a presenter in a brightly-coloured set
Rugby	Originally 16x9 component, but converted to 4x3	Rugby match. This clip could not consistently achieve *quantizer_scale*=6 as too much bit-rate was required.
Top Gear	Originally 16x9 component, but converted to 4x3	Motoring magazine. Clip shows Quentin Willson standing in front of a fairground ride.
Top Of The Pops	PAL 4x3	Pop Music show. Clip shows a crowd watching Bus Stop & Carl Douglas.
Wipeout	PAL 4x3	Quiz show. Clip shows Bob Monkhouse talking and gesticulating.

screen). These programmes were then requested on tape from BBC Television Centre. The remaining items used were taken from a test tape that had been put together to test MPEG-2 coders. The intention was to show a wide range of programmes, representative of broadcast output, and which showed a range of sensitivities to quantization artefacts. The clips were collected together, the aspect-ratio converted where necessary, and they were then recorded onto Digital Betacam tape.

These programme items were then recorded through the coder (onto Digital Betacam tape) and edited into three short tests, which were shown to a small number of viewers to get an idea as to the range of quality which should be included in the main tests.

The programme items used

These are shown in *Table 1*.

The test method

The test method used is known as the DSCQS test and is described fully in ITU-R Recommendation BT.500 [1].

Each individual test consists of two presentations, A and B, which always originate from the same source clip, but one is coded and the other is the uncoded “reference” picture. The viewers are asked to grade both pictures, and are not told which is the reference picture. The position of the reference picture varies according to a pseudo-random sequence. The test tape was made up from an EDL, generated from a specially-written computer program. Viewers see each presentation twice (A,B,A,B), according to the test format shown in *Table 2*.

Table 2
The test format.

Item	Duration (seconds)
Test number (white on grey background)	2
Presentation A	8 - 10
Pause (grey screen)	3
Presentation B	8 - 10
Pause (grey screen)	3
Presentation A again	8 - 10
Pause (grey screen)	3
Presentation B again	8 - 10
Pause to vote (grey screen)	8

A test block contained 22 tests, and lasted about 20 minutes. There were three blocks in all, and the order in which viewers watched the blocks was changed from session to session. The viewers used a scale like the one in *Fig. 1* to mark their results.

The viewers marked the scale in blue ink with a horizontal line to represent their opinion of the quality of a particular presentation.

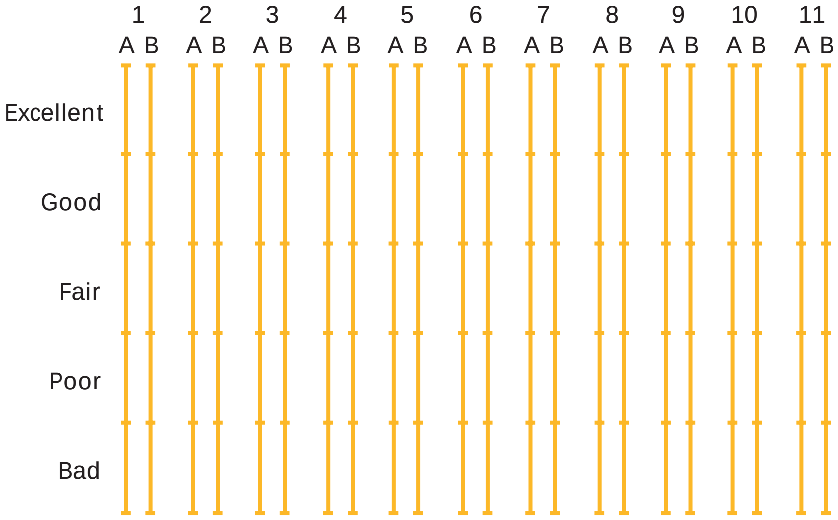


Figure 1
Part of the score sheet used by the viewers.

The data from the score sheets was then entered into a spreadsheet for analyzis. The figure entered into the computer was the difference in millimetres between presentations A and B, and the number was negative if the viewer had graded the coded picture as being worse.

The viewing room

To allow the test to be presented to three viewers at once (saving time), the room was divided into three viewing booths using mobile screens. These were arranged in an arc along one wall (see *Fig. 2*). Each booth was provided with a monitor on appropriate staging, and a chair at 4H from the screen. On top of the monitor was a display which lit up to show either "A" or "B", to remind the viewer which test presentation they were watching.

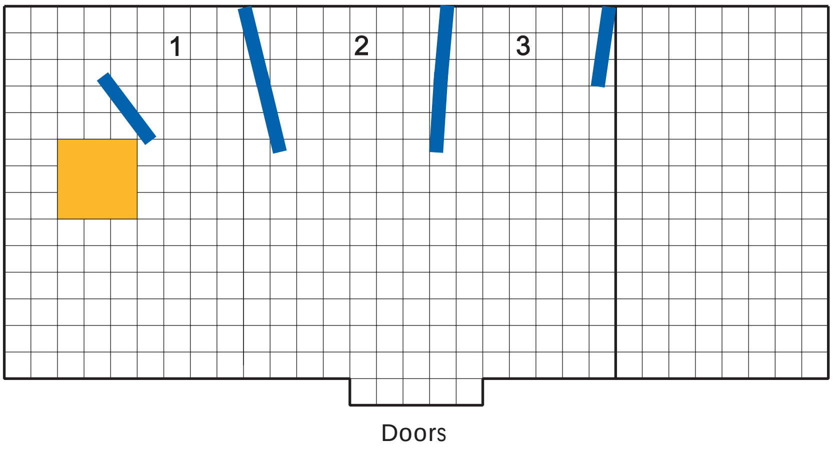


Figure 2
Floor plan of viewing room used for the tests: one square represents 30 cm^2 approx.

The monitors (Ikegami 19" digital input) were all set up and matched before the test sessions began, using their built-in pattern generators. The test tape was played on a Sony Digital Betacam VT machine (in the Record Suite) which was controlled remotely by the experimenter. The A/B displays were operated by tones recorded on the audio tracks of the test tapes.

The fluorescent lighting in the room was dimmed to a low setting, and the windows were blacked out with card and “gaffer” tape [3]. The lighting and other viewing conditions were set up in accordance with the guidelines given in ITU-R Recommendation BT.500 [1].

The training procedure

To help the viewers make an informed judgement during the tests, a training session was given. This was divided into two parts. In the first part, the viewers were invited to look closely at a monitor whilst a number of video clips were played. The uncoded clip was shown first, and then the coded clip was shown three times. The clips shown ranged from very coarse to very fine quantization, to give an idea of the range of quality which would be seen in the tests.

The second section of training was a short test, designed to introduce viewers to the test format. The results of this test were discarded.

The viewers

Twenty-six viewers performed the tests over a two-week period. They were all drawn from BBC R&D staff and covered a wide range of expertise. Of the viewers, there were at least nine “professional” viewers (people whose work brought them into direct contact with TV pictures on a regular basis) and the remainder were made up from researchers in other fields, office staff and trainees. They all volunteered (or were volunteered!) to take the tests.

Quantization settings

The base *quantizer_scale* was controlled. A dynamic offset (positive or negative) was added by the coder’s adaptive quantization algorithm on a macroblock-by-macroblock basis; this helped to make the subjective quality less dependent on picture content. The *quantizer_scale* parameter is defined in the MPEG video coding standards, ISO/IEC 11172-2 and ISO/IEC 13818-2 [2]. Note that the results refer to *quantizer_scale* and not *quantizer_scale_ code*.

[3] Cloth-based tape used for marking camera positions on the studio floor, for example.

The values quoted in the results are for the base *quantizer_scale*. The test used settings of 6, 8, 10, 12, 14 and 16 for each programme item (based on the results of the preliminary tests). *Appendix 1* gives the approximate average bit-rates for the test items.

Abbreviations

DSCQS	Double-stimulus continuous quality scale
DCT	Discrete cosine transform
EDL	Edit decision list
HD	High definition
MPEG	Moving Picture Experts Group
PAL	Phase alternation line

Discussion of results

The plots in *Appendix 2* show the mean score recorded by the viewers for each setting of *quantizer_scale*. The scores are directly related to the 100 millimetre scale on the graph sheet, so a score of – 20 means that the viewers, on average, thought that the coded picture was 20 mm worse than the reference. On this scale, 20 mm corresponds to approximately one ITU-R grade, so a score of – 20 would mean that an “excellent” reference picture had been reduced to merely “good” by the coding process at that level of quantization.

Individual plots for seven of the programme items are shown in *Figs. A3 - A9*. On these plots, the vertical bars indicate both the standard deviation and the 95% confidence interval. The standard deviation shown is the actual standard deviation of the viewers’ scores (Microsoft Excel’s STDEVP function). This is intended to give a measure of the range of responses received for a given test. (For normally distributed data, 68% of responses lie within ± 1 standard deviation of the mean).

Because only a sample of the viewing population was taken (i.e. the 26 people who took part in the tests), the 95% confidence interval is used to show the accuracy of the sampling. We can be 95% certain that the actual mean plot (i.e. that for the entire UK viewing population) will lie between the error bars shown.

It was not found necessary to exclude any of the viewers from the results – only one viewer was identified as not fitting the general trend of the others, and removing him was found to have no significant effect on the rest of the results.

Looking now at the results in general – in particular the plot *(Fig. A1)* showing all of the programme items – the relationship between *quantizer_scale* and mean score appears to be largely linear, although the rate of degradation varies a great deal between the programme items. We did not observe any clear “threshold of visibility” – a sharp drop in picture quality at a certain level of quantization.

It can be seen that it is difficult to control the picture quality in a programme-independent manner, although it would seem that this method (fixed *quantizer_scale*) was more reliable than by simply fixing the bit-rate. The variations between programmes are probably introduced by (i) variations in the picture quality of the source (some programme items had been PAL-coded and some had film noise) and (ii) the general variations in picture content and motion. It is also worth noting that the 4:2:2 to 4:2:0 chrominance subsampling affected the picture quality slightly, although the majority of the viewers did not comment on any chrominance effects.

Conclusions

In most cases, setting a minimum slice quantization of 6 introduced a quality degradation of less than half a grade on the ITU-R quality scale, when compared with the source picture. With the slice quantization set to 8, the quality degradation introduced is (nearly always) less than one grade. However, there is considerable variation between programmes: a one-grade degradation (a score of – 20) is achieved at *quantizer_scale* = 8 for *"Top Gear"* and *"GW Garden"*, whereas *"Top of the Pops"* does not achieve this level until *quantizer_scale* = 14. *"Ironside"* and *"Heart By-pass"* don't even get as low as – 20, even at the coarsest quantization tested.

Limiting the *quantizer_scale* (e.g. to a minimum of 8) – in order to provide an opportunistic data service – would reduce the picture quality on some programmes (BBC News 24's studio shots are currently coded at *quantizer_scale* = 4 and BBC 1's *"Teletubbies"* varies between 5 and 8), but since many other programmes are transmitted at much coarser quantizations (the rolling trailers on BBC Choice and BBC Knowledge are a case in point), such a limit would make the picture quality more consistent. In the end, a decision has to be made as to whether the data capacity provided is worth the (relatively small) degradation in picture quality. It should be noted that buffer regulation to prevent decoder over- or under-flow is non-trivial for variable bit-rate coding.

Further work to support this could involve:

- tests with other coders to see if similar results are observed;

John Fletcher studied Physics and Electrical Sciences at the University of Cambridge, gaining a Master of Arts degree. He joined the BBC's Engineering Research Department in 1986.

During his early career, he worked in the field of audio and acoustics. More recently, his research interests have focussed on MPEG video coding; particularly with regard to making efficient use of the available transmission bandwidth.

Mr Fletcher is currently spending six months in BBC Production as a Technical Advisor.

Michael Prior-Jones joined the BBC's Research & Development department in 1998 as a trainee engineer. He worked on data-logging systems for RF spectrum planning and on research into MPEG-2 coding. He is now on a four-year M.Eng. degree course in Electronic Engineering at the University of York and hopes to be returning to the BBC during university vacations.

- ⇨ measurements of quantization on programmes as currently broadcast, to estimate the number and genres of programmes that would be affected by the introduction of an opportunistic data service;
- ⇨ an investigation into the bit-rate required for a given quantization for a variety of programme items.

Bibliography

[1] ITU-R Recommendation BT.500: **Methodology for the subjective assessment of the quality of television pictures**
ITU, Geneva, 1998.
http://www.itu.int/itudoc/itu-r/rec/bt/500-9.html

[2] ISO/IEC 13818-2: **Information technology – Generic coding of moving pictures and associated audio information: Video**
ISO/IEC, Geneva,1996.
http://www.iso.ch/cate/d22990.html#0

Appendix 1
Approximate average bit-rates (Mbit/s) of the test items

Clips	quantizer_scale					
	6	8	10	12	14	16
Eastenders	12	6	5	3	2. 5	2
Wipeout	7	4	3	2	1. 5	1. 5
Playdays	12	6	5. 5	5	3	2
Top of the Pops	13	10	8	7	5	5
Heart By-pass	9	7	5	4	3	3
Ironside	13	6	5	4	3. 5	3
Top Gear	13	9	7	6	5	4
GW Potting Up	11	6	4	3	3	3
GW Garden	12	8	6	4	3	3
Rugby	15	10	8	6	5	4
Box Hill	9	5	4	3. 5	3	2

Appendix 2
Graphs

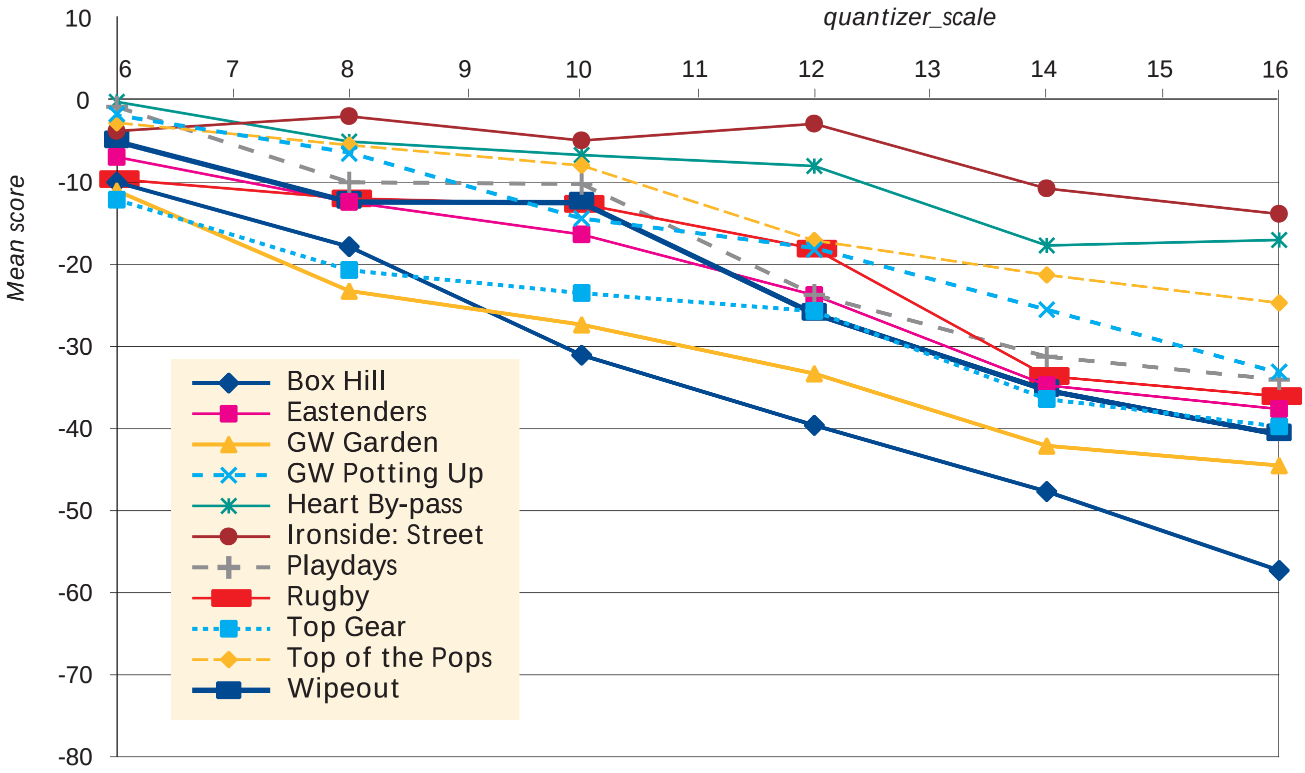


Figure A1
Plot showing all programme items.

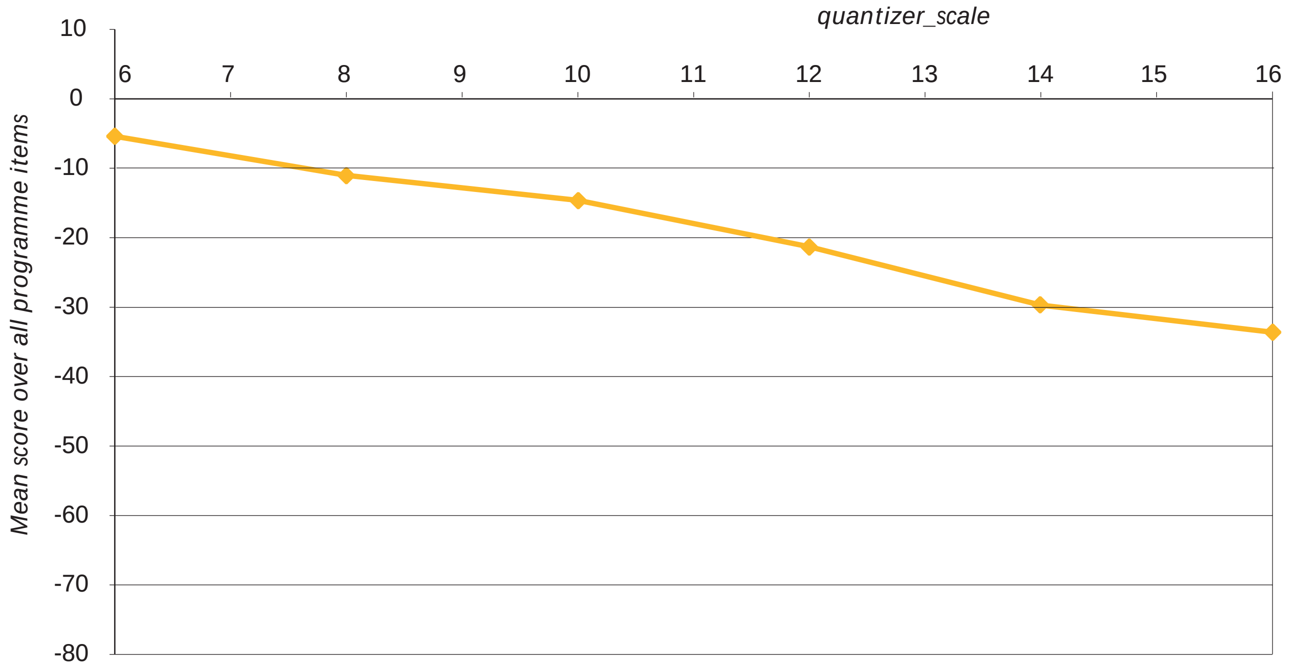


Figure A2
Overall plot (averaged over all programme items).

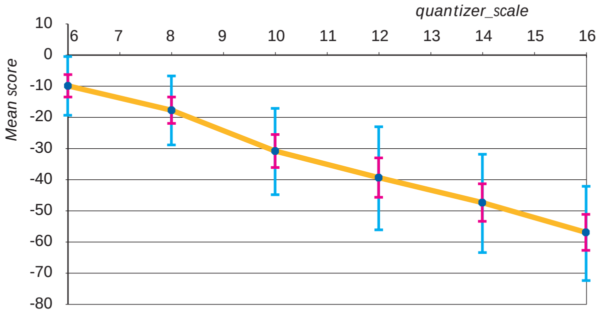


Figure A3
Box Hill.

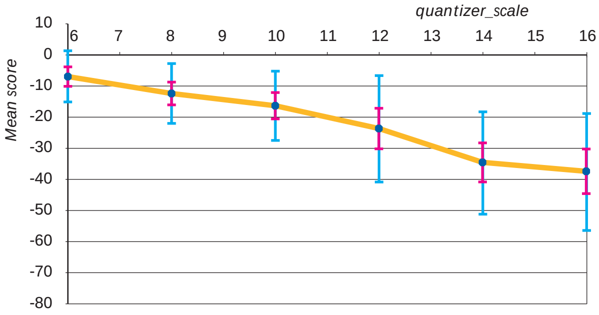


Figure A4
Eastenders.

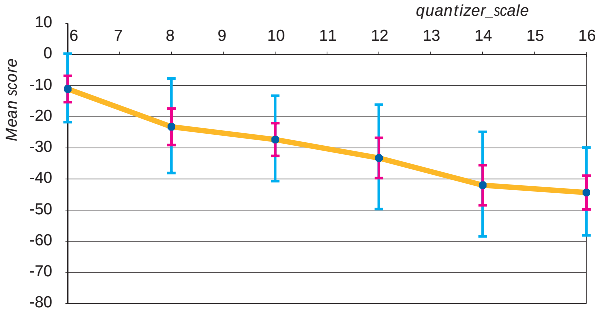


Figure A5
GW Garden.

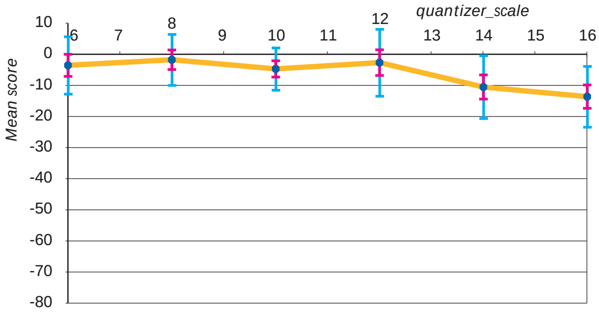


Figure A6
Ironside.

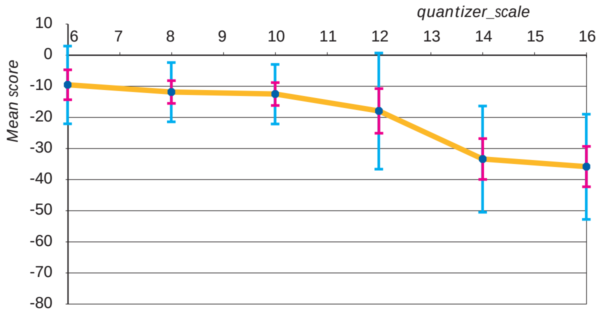


Figure A7
Rugby.

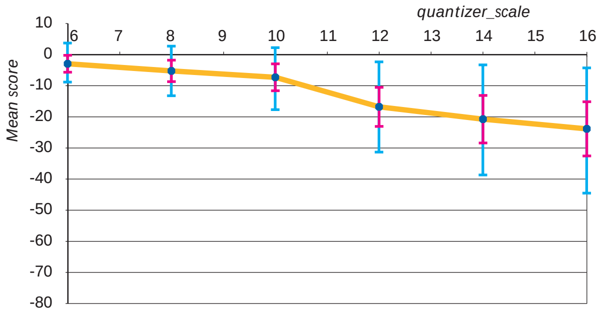


Figure A8
Top of the Pops.

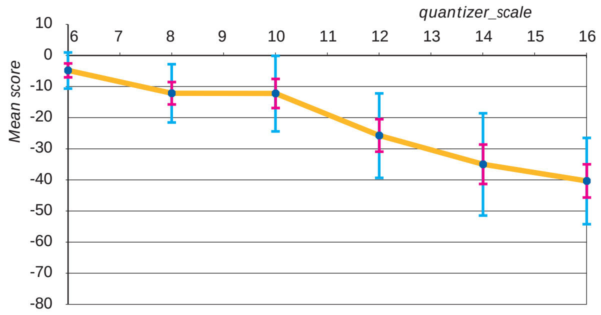
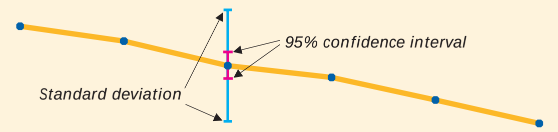


Figure A9
Wipeout.

Explanation of error bars.


95% confidence interval
Standard deviation