

# National curriculum test reviews

## Trends over time: 2000–7



February 2009  
Ofqual/09/3982

Paul E Newton  
Assessment research team

## Contents

Rationale .....	3
How valid are the data? .....	4
How robust are the data? .....	5
The situation prior to 2004 .....	5
The situation in 2004 .....	6
The situation in 2008 .....	6
What can be inferred from the data? .....	8
The number of students for whom reviews were requested .....	9
The number of students whose review requests were successful .....	9
The percentage of successful review outcomes for which students' levels were lowered .....	10
Conclusions .....	12
References .....	13
Number of review requests – mathematics and science .....	14
Number of review requests – English .....	15
Number of successful requests – mathematics and science .....	16
Number of successful requests – English .....	17
Percentage change over time – mathematics and science .....	18
Percentage change over time – English .....	19
Percentage of successful reviews taking levels down .....	20
Percentage of requests that were successful .....	21

## Rationale

On 17 November 2004, the Qualifications and Curriculum Authority (QCA) published a review of the delivery of the 2004 key stage 3 English tests, following delays and operational difficulties experienced by schools that year. A Committee of Review was established by the QCA Board in September, and part of its remit was to consider whether the quality of marking had suffered. While the committee found no reason to believe that the test itself, the marking quality or the final national results were in doubt, it expressed concern at the available evidence. The committee concluded:

The criterion for judging marking quality normally used is the number of reviews submitted by schools that are upheld by the External Marking Agency (AQA), expressed as a percentage of the whole cohort. The review team found this an unreliable and unsatisfactory criterion due to the inconsistencies in the way data has been collected from year to year and the lack of information on the reasons why schools may or may not decide to request a review.

This report examines trends over time in data from the national curriculum test review process from 2000 to 2007. It considers what inferences, if any, can be drawn from them regarding quality of marking over time.

## How valid are the data?

Even in the best of circumstances, review data represent an invalid index of marking quality, since schools are not required to make reviews and are likely to be more or less inclined to request them from one year to the next for reasons that have nothing to do with the quality of marking of their students' scripts, such as:

- a delay in scripts being returned to schools (may decrease the likelihood of review requests if, for example, the time available for checking scripts has lapsed)
- a lot of negative media reporting (may increase the likelihood of review requests if, for example, schools are more attuned to the possibility of marking error and check scripts more diligently).

A survey of schools in 2000 revealed considerable differences in the extent to which schools checked scripts for marking error, and considerable differences in schools' attitudes to requesting reviews (Newton and Whetton, 2000, 2005). Evidence also shows that attitudes change over time (Newton, 2004).

In short, even in the best of circumstances, evidence from the national curriculum test review process could not be assumed to constitute a valid index of trends in marking quality over time. Of course, the data were never designed to be used for this purpose, and their use for this purpose has been largely incidental.

## How robust are the data?

The following three sections consider whether the state of the data represents the 'best of circumstances' or whether a lack of robustness presents further reason to question the validity of evidence from the review process.

### The situation prior to 2004

As long ago as 2000, when there were three external marking agencies (EMAs) working independently, serious problems were noted for the interpretation of data from the review process (Newton and Whetton, 2000).

First, in 2000, there was no clear ruling on how the possibility of separate reviews for (key stage 2) reading, writing and English overall should be handled; and evidence showed that the EMAs may have adopted different stances. For instance, if a school submitted a review request for writing that changed both the writing level and the English level, this could be classified in a number of ways, including:

- as one review request (W) and one level change (W) – assuming that writing was the basis for the request
- as one review request (E) and one level change (E) – assuming that English is the higher level unit for reporting purposes
- as one review request (W or E) and two level changes (W, E)
- as two review requests (W, E) and two level changes (W, E).

Second, there was ambiguity in the interpretation of the number of students involved in group reviews (GRs), which, incidentally, were only available for key stage 3 English at that point in time. It was not clear whether the value ought to reflect:

- the total number of pupils in the group that was submitted for review
- the total number of pupils in the school cohort who took the key stage 3 English test.

Again, in 2000, the EMAs took different approaches. One was only able to record the number of students whose levels actually changed.

Third, the review databases contained errors that could easily have been avoided through better data management, which casts further doubt on the robustness of the information gathered.

## The situation in 2004

In a review of trends over time produced for the Committee of Review (Newton, 2004) a number of concerns with the data were noted, including the following.

First, from 1999 it was possible to request reviews for reading and writing separately at key stage 2, which would have inflated the number of review requests and successful outcomes in comparison with previous years. However, as mentioned above, this inflation may not have been recorded consistently across EMAs.

In 2003, when levels were first awarded separately for reading and writing on the key stage 3 English tests, separate reviews for the components were not allowed.

Second, general doubt was expressed over the validity of the figures for key stage 2 English, both in 2000 and in subsequent years (a view reinforced by the observation that both QCA and EMA officers responsible for processing reviews data had come and gone with some frequency over the years).

Third, some of the published data were wrong. For example, the 1998 review request data for key stage 3 mathematics were not available, due to errors in the Standards report. Incorrect figures were published in 1998; in fact, the report simply repeated the figures for science. Similarly, the figure published in the 1997 Standards report for key stage 2 science marking reviews (664) was wrong (and the revised figure of 1,808 was published in the subsequent year).

## The situation in 2008

Despite the possibility of requesting reviews for reading and writing separately, data have only ever been published at the overall level for English. How the level changes for reading and writing have been recorded over time – most importantly, whether a consistent approach has been adopted – is not clear. When figures were discussed with the department in 2004, different approaches to their calculation had been taken, which led to some confusion (Newton, 2004). How these data have been collated subsequently is also unclear.

Repeating a trend over time, more errors in the presentation of review data were found in the reports published since 2004. For example:

- in the published clerical review data for key stage 3 English in 2004, the total number of reviews requested was equal to the number of level changes up and down (that is, all claimed level changes appeared to have been granted) – this seems likely to have been a transcription error of some kind – the data for 2004 were inconsistent with both previous and subsequent years
- transcription errors were present in the published review data for key stage 2 English, mathematics and science in 2007, as well as for key stage 3 English –

these typically involved misrepresenting the percentage of the national cohort involved in review requests by a factor of 10 (for example 0.52 per cent of the cohort when the correct figure was 0.052 per cent)

- an investigation into further apparent inconsistencies for the 2007 data revealed problems with the data published in 2006, particularly in relation to the numbers of students for whom reviews were requested.

While the transcription errors in 2007 were serious, of more concern was the potential corruption of the data for both 2006 and 2007. Many of the published figures (based on marking agency management information) were at odds with data collected by the QCA (the National Assessment Agency at the time) principal data analyst (based on separate data feeds). Since the analyst collected review data consistently from 2005 to 2007, it was decided to base the following analyses upon:

- published review data – published annually within technical appendices to Standards reports (later published separately) – from 2000 to 2004
- previously unpublished review data – collated annually by the QCA principal data analyst – from 2005 to 2007.<sup>1</sup>

The published and collated data do not differ for 2005, only for 2006 and 2007 (which is further reason to rely upon the collated data for the following analyses). For 2006 and 2007 they tend not to differ radically in terms of review outcomes, but they differ sometimes very radically in terms of review requests. The collated data may not necessarily be 'correct', but at least they seem to be consistent.

---

<sup>1</sup> QCA has recently decided to publish the data collated by the principal data analyst, from 2005 to 2007, in correction of data published previously.

## What can be inferred from the data?

Apparent trends over time have been presented graphically in figures 1 to 32, presented at the end of this report.

Five types of trend were analysed:

- the number of students for whom reviews were requested
- the number of students whose review requests were successful (that is, the number of requests that resulted in an upward or downward change of level)
- the percentage change in successful review outcomes from year to year (that is, the change from one year to the next in the number of requests that resulted in an upward or downward change of level)
- the percentage of successful review outcomes for which students' levels were lowered
- the percentage of review requests that were successful.

From the outset, it is worth noting that review outcome data are traditionally published as both raw numbers and percentages of the national cohort. Were the cohort size to change from one year to the next, the comparison of raw numbers would be at least somewhat misleading. In practice, the cohort size does change, although the impact on the percentages would not be very large. More significantly, though, the percentage figures in the published review data (prior to 2008) were based on a constant cohort of 650,000. As such, the trend lines, based on raw numbers and percentages, would be identical.

For most of the following analyses, the data for English (E) have been separated from mathematics and science (M/S), for the simple reason that the figures for English tend to be much higher and compress the scale for the other subjects. It is important to consider the scale against which the lines are judged (the y-axis values) when comparing patterns between E and M/S.

The data were broken down as follows:

- R1 review – a clerical check – are the marks added correctly (etc)?
- R2 review – a check of marking – has the mark scheme been applied correctly for an individual student?
- GR – a group review – has the mark scheme been applied correctly for a group of students?

- ALL – all three types of review combined.

### **The number of students for whom reviews were requested**

The number of students for whom R1 reviews were requested seems to be fairly stable over time, for M/S (figure 1), as well as for English (figure 5). There is one obvious glitch for 3E, which is probably a transcription error of some kind, as noted above.

The number of students for whom R2 reviews were requested has been somewhat less stable for 3S, 2E and 3E over time (figures 2 and 6). Both 3E and 3S show a peak in 2004, from which requests appear to have fallen.

In fact, trends in patterns for R2 reviews are harder to interpret when the possibility of a GR exists. For example, the apparent drop in R2 requests for 3S from 2004 to 2005 was counter-balanced (to some extent) by an opportunity for GR reviews in 2005. A similar counter-balancing seems to have occurred for 3E.

As it happens, the trend for 3E is hard to map because no data exist on the number of requests for students involved in GR reviews from 2000 to 2004 (figure 7). This is probably for reasons mentioned earlier, to do with how this value is best defined. The same lack of data occurs for 2E in 2004, too, and explains the lack of data points on the ALL reviews graphs (figure 8).<sup>2</sup>

### **The number of students whose review requests were successful**

The number of students (as a percentage of the cohort) whose review requests were successful – either up or down – tends to be taken as the most direct estimate of marking quality. If fewer level changes are made, then, all other things being equal, the marking must have been more reliable.

Generally speaking, the number of students for whom successful R1 review requests were made has declined over time, particularly from 2004 (figures 9 and 13). The one oddity, again, is the glitch for 3E in 2004. If the number of successful reviews declines, while the number of review requests remains fairly stable, this suggests that the percentage of review requests that are successful must decline correspondingly.

Inspection of figure 29 shows a radical decrease in the percentage of students for whom an R1 review request was successful. So what exactly is supposed to be happening here? From similar computations underlying a previous review (Newton, 2004) R1 review requests were consistently successful, from 1999 to 2003, in the

---

<sup>2</sup> Note that GR has been permitted for 3E from prior to 2000, but has only been permitted for 2E from 2004.

region of 80 per cent to 95 per cent. The one notable exception is a drop in 2003 for 3E. The only exception to this pattern was 2E for which R1 review requests were consistently successful, from 1999 to 2003, but in the lower region of 45 per cent to 60 per cent. In stark contrast, figure 29 suggests a radical drop, since 2004, to less than 25 per cent success for four of the six tests (see 2006 in particular).

Remember that an R1 review is simply a clerical check. It should be transparent whether or not such errors have occurred. There would seem to be a number of possible alternative explanations, either:

- schools really have become far worse at identifying genuine clerical errors
- procedures for conducting R1 reviews have changed radically
- the data are being collated differently and/or incorrectly.

The latter might seem to be most likely. If so, then this casts considerable doubt on the data since 2004. Unfortunately, exactly which aspects of the data are thrown into question is not clear.

Turning back to figure 10, the (raw) number of successful R2 reviews has remained fairly stable over time, despite a slight jump for 3S in 2004. For 2E and 3E (figure 14) the figures seem to fall slightly more than for M/S.

In fact, the picture is considerably more complicated for 3E, since the number of successful GRs seems to have plummeted from 13,390 in 2004 to only 3,071 in 2005 (figure 15). Given the lack of data prior to 2005 on the number of students involved in GR requests, it is not clear how to account for this drop.

To explore the extent of change that might be anticipated from one year to the next, tables were created that expressed the change in the number of successful review outcomes, from one year to the next, as a percentage of the preceding year's value. So, a change from 320 level changes in year 1 to 640 level changes in year 2 would be an increase of 100 per cent. These data are presented graphically in figures 17 to 24. The data are little more than a re-description of the trends presented in figures 9 to 16, so will not be discussed further.

### **The percentage of successful review outcomes for which students' levels were lowered**

One final point worthy of mention derives from figure 25. This captures a trend in the proportion of level changes that lowered, rather than raised, students' levels. A change in this value is, presumably, an indication of trends in the extent to which schools are inclined to request reviews that lower students' levels rather than raise them. In theory, there ought to be as many requests for reviews to lower students'

levels as to raise them, particularly for clerical error (R1). There is no obvious reason why such errors ought to be biased against students.

Interestingly, although 'raising' reviews were in the clear majority in the early days of national curriculum testing, there was a clear trend – from 2000 to 2002 across all subjects – for the percentage of 'lowering' reviews to increase. Even more interestingly, though, this trend reversed between 2002 and 2007 across all subjects. This suggests that schools are now less inclined than ever to request reviews to lower students' levels.

## Conclusions

The foregoing analyses presented data from 2000 to 2007, providing a context for the interpretation of figures from 2008.

There are strong reasons to doubt the robustness of data on review requests and outcomes from 2000 to 2007, although it is not clear exactly where the problems might lie. There is certainly good reason to question the consistency of data over time, meaning that extreme caution should be exercised when drawing any inferences from the trend lines.

Even in the best of circumstances, data from the review process would be invalid as an index of change in marking quality, since there are all sorts of reasons why schools request reviews, many of which are independent of the quality of marking of their scripts. The change over time in the percentage of 'lowering' R1 reviews is a prime example of this.

It needs also to be recalled that the process for 2008 was very different from that in previous years, with the appointment of a new contractor as well as the subsequent transferral of responsibility for processing reviews to the QCA (NAA). Whether data can be produced in a format that is consistent with previous years remains to be seen.

In summary, the negative conclusion of the 2004 Committee of Review holds true in 2008: review data represent an 'unreliable and unsatisfactory criterion due to the inconsistencies in the way data has been collected from year to year and the lack of information on the reasons why schools may or may not decide to request a review'.

## References

Committee of Review. *Report on key stage 3 English review of service delivery failure 2003–2004 to QCA Board*. London: Qualifications and Curriculum Authority, 2004. ([www.qca.org.uk/libraryAssets/media/10343\\_ks3\\_en\\_report\\_04.pdf](http://www.qca.org.uk/libraryAssets/media/10343_ks3_en_report_04.pdf)).

Newton, Paul E, and Whetton, Chris. *An evaluation of external marking review services during 2000*. London: Qualifications and Curriculum Authority, 2000.

Newton, Paul E, and Whetton, Chris. 'The effectiveness of systems for appealing against marking error'. *Oxford Review of Education*. 31.2 (2005): 273–291.

Newton, Paul E. *Data on national curriculum review requests, 1995 to 2003*. Unpublished QCA Board Committee of Review Paper, 2004.

Fig. 1. R1 - No. Review Requests

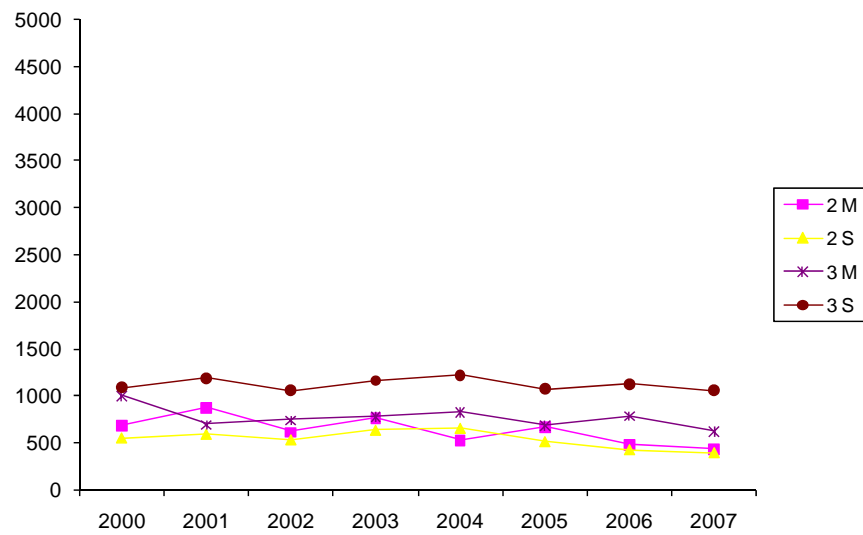


Fig. 2. R2 - No. Review Requests

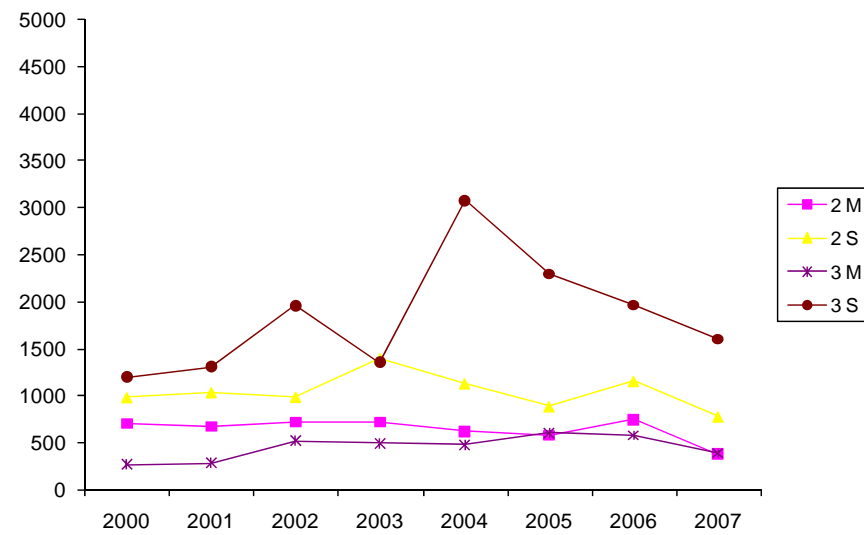


Fig. 3. GR - No. Review Requests

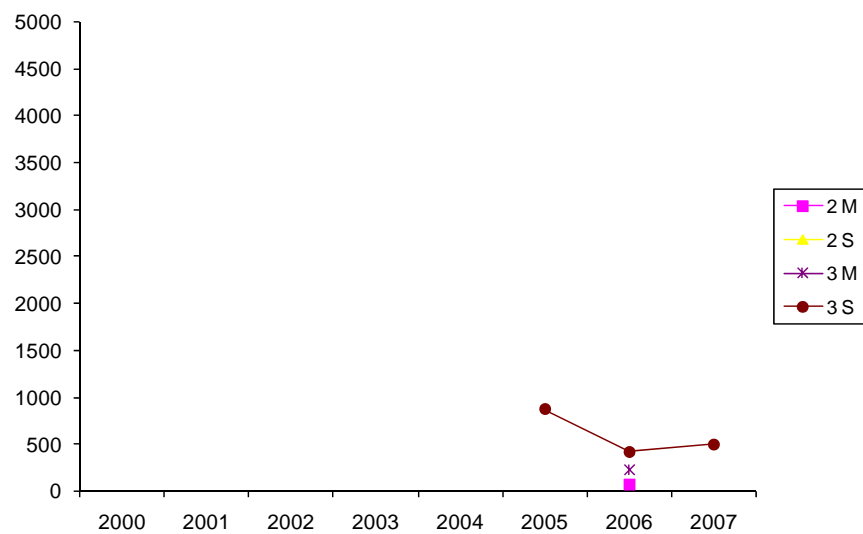
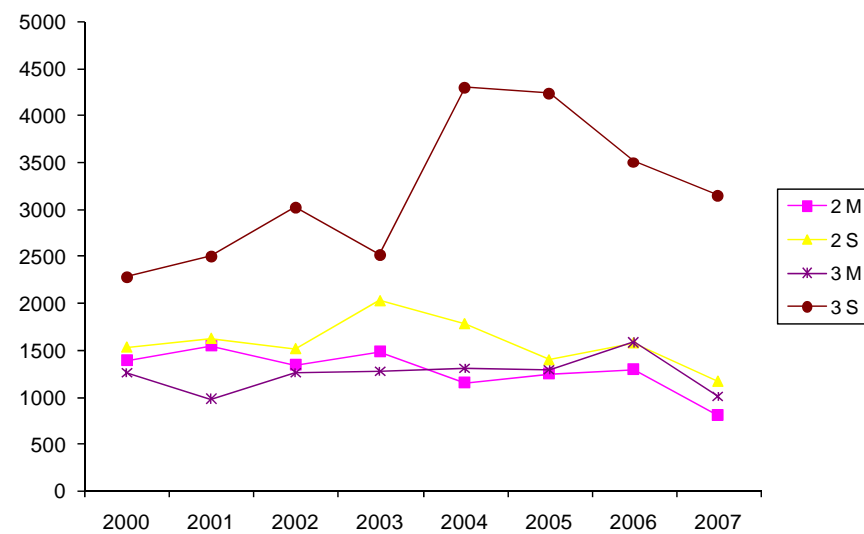
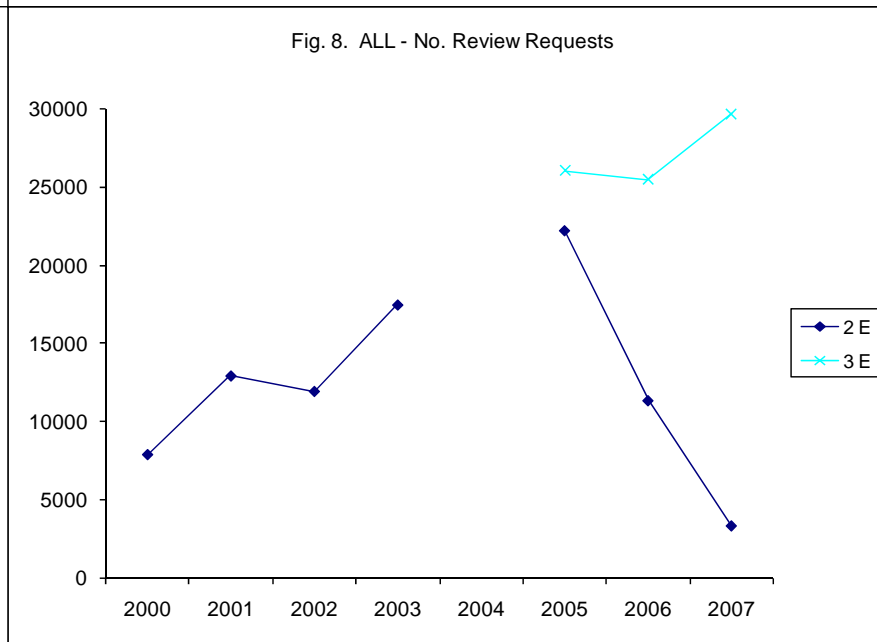
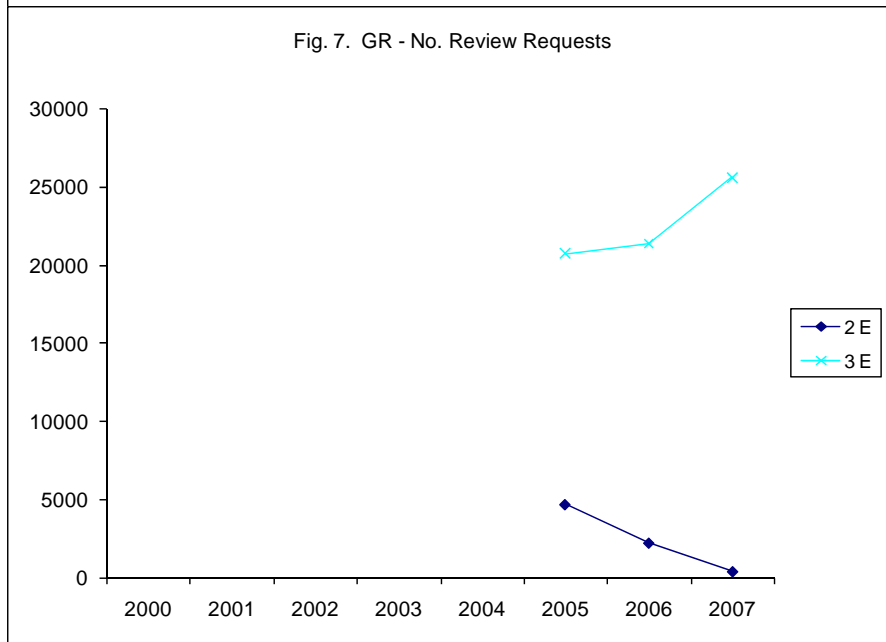
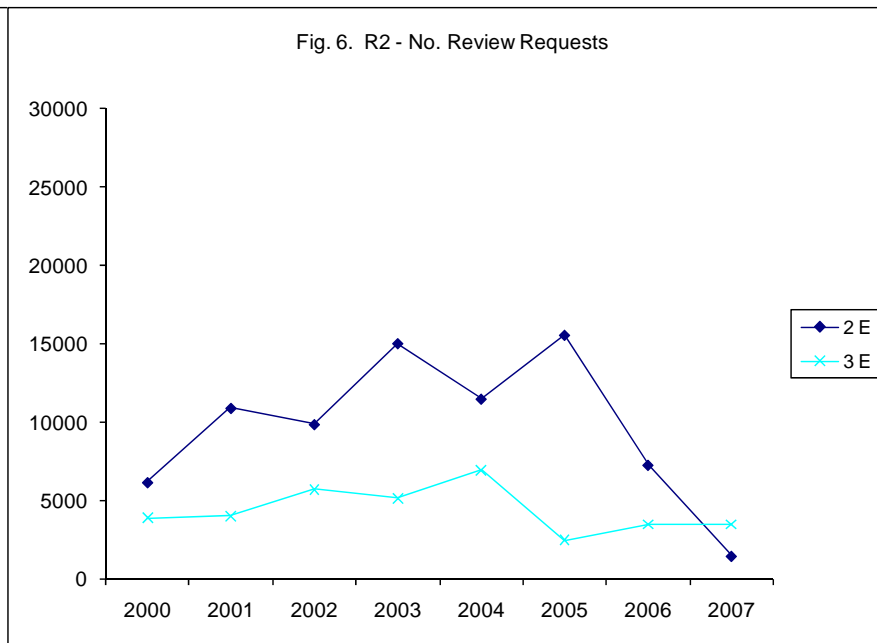
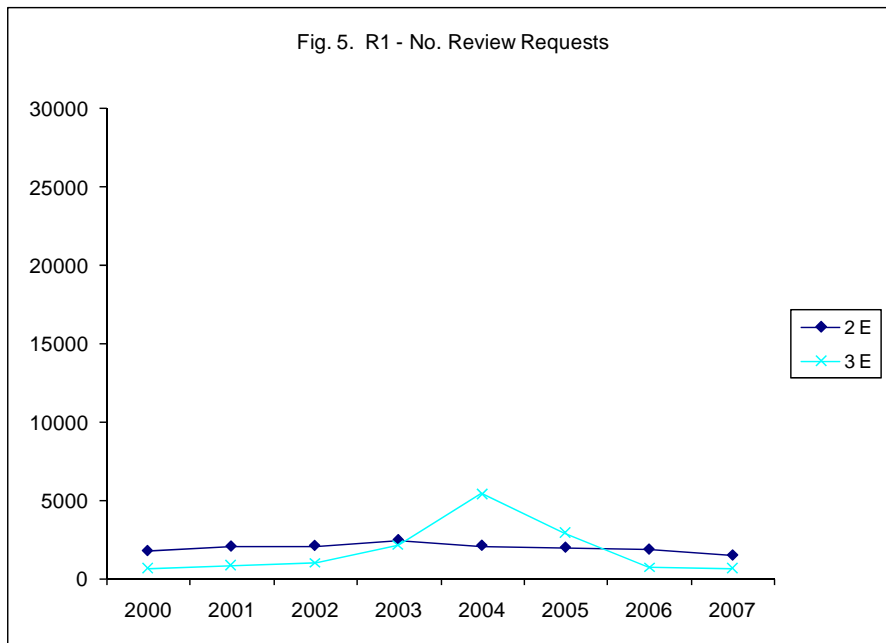
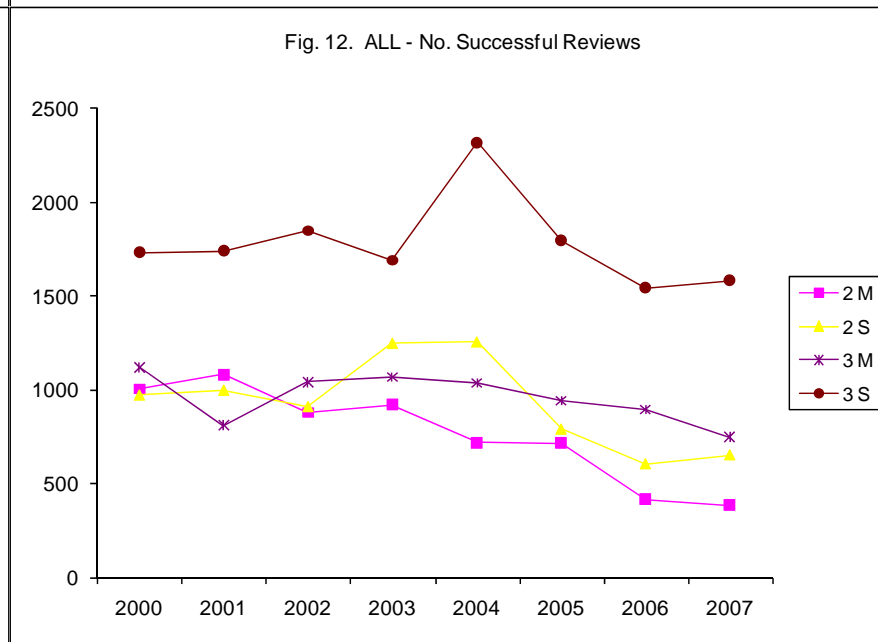
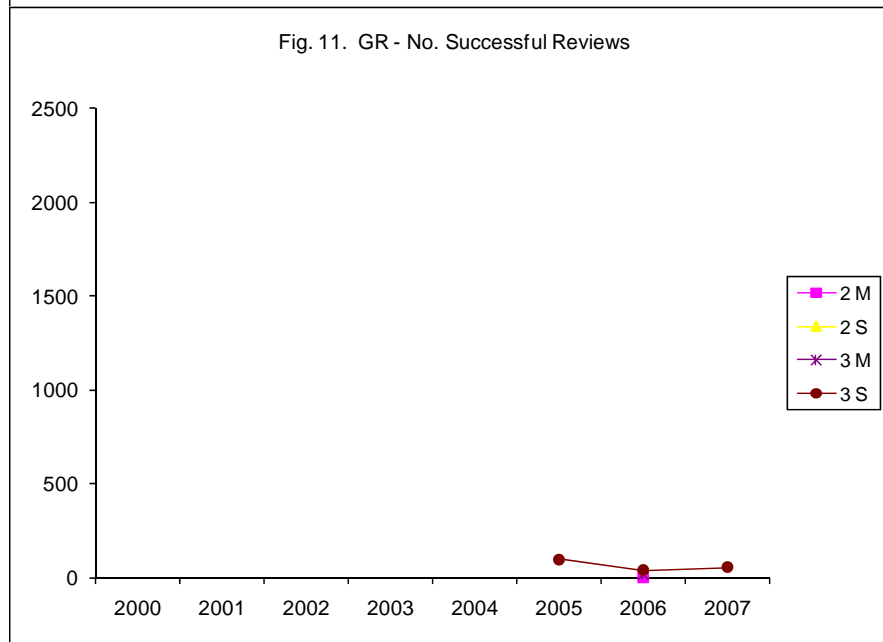
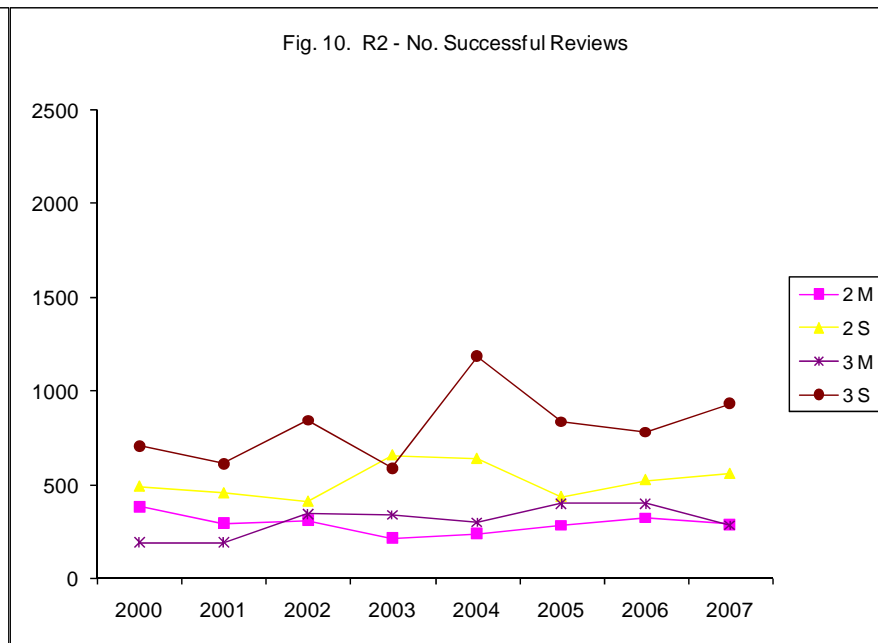
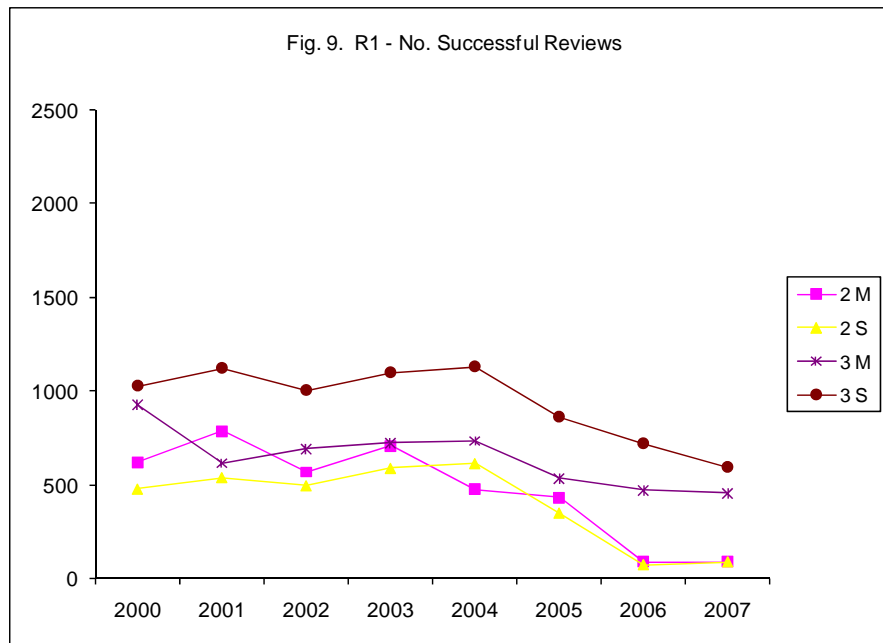
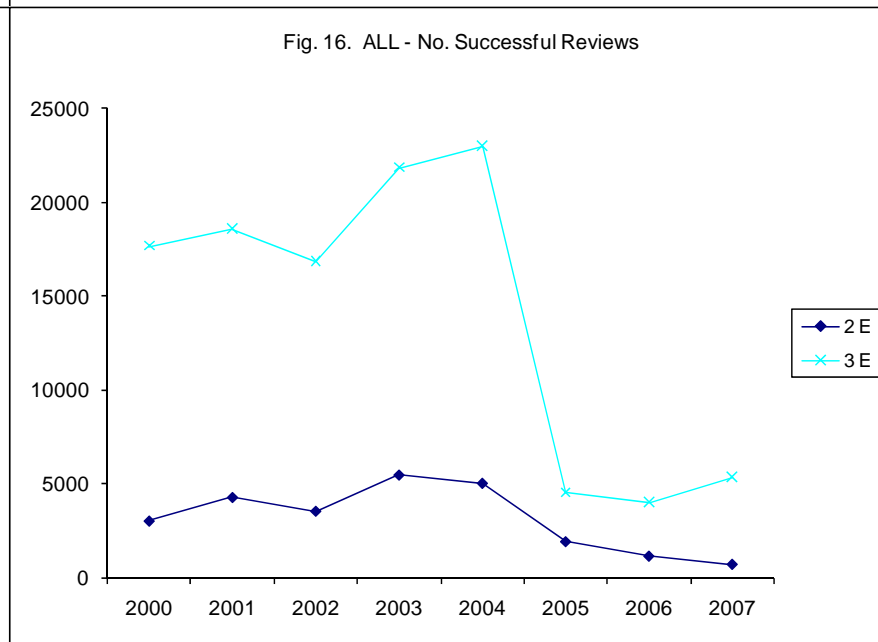
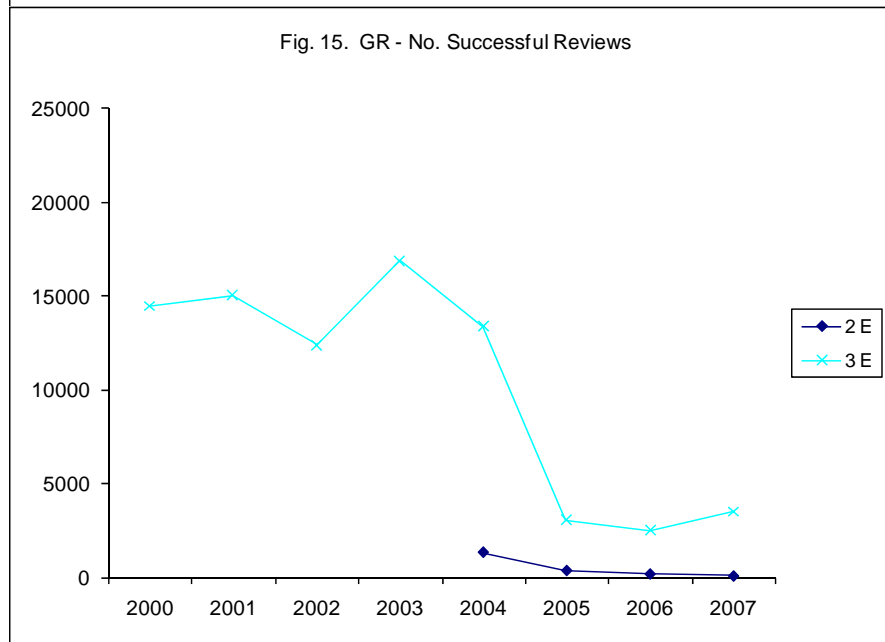
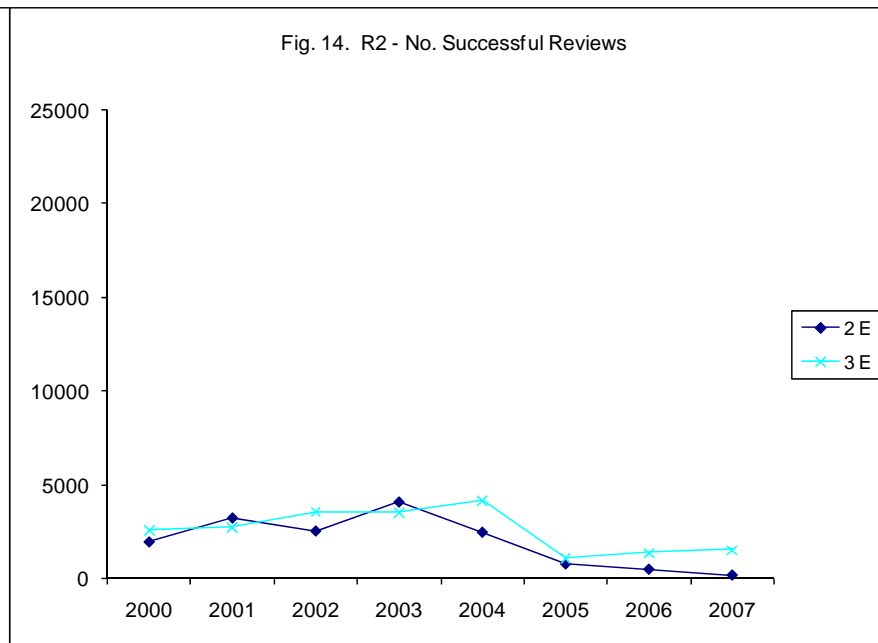
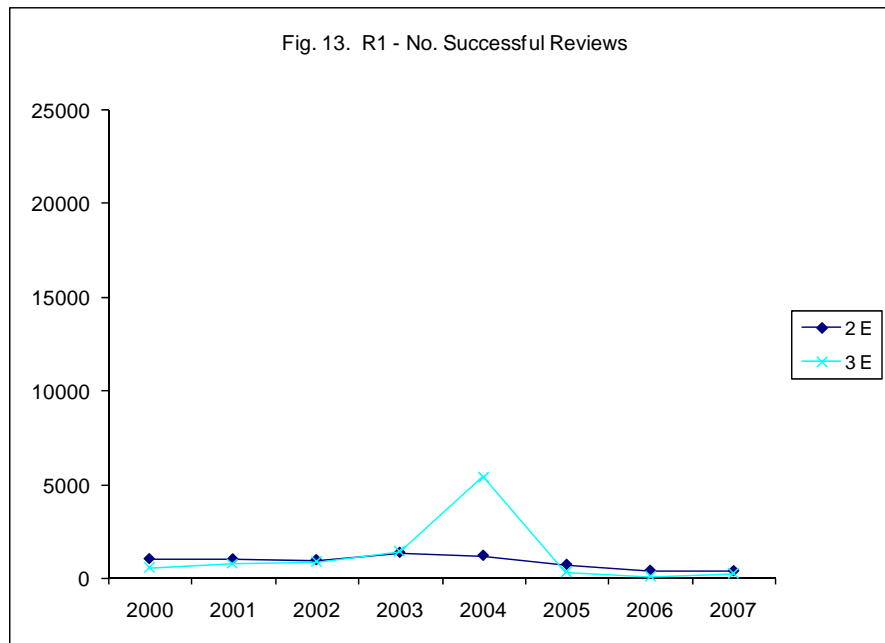


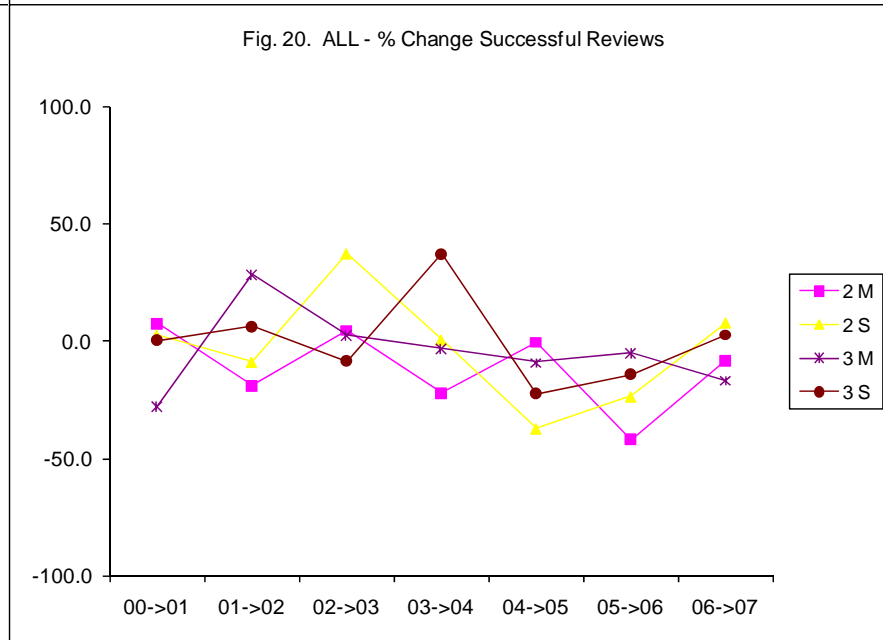
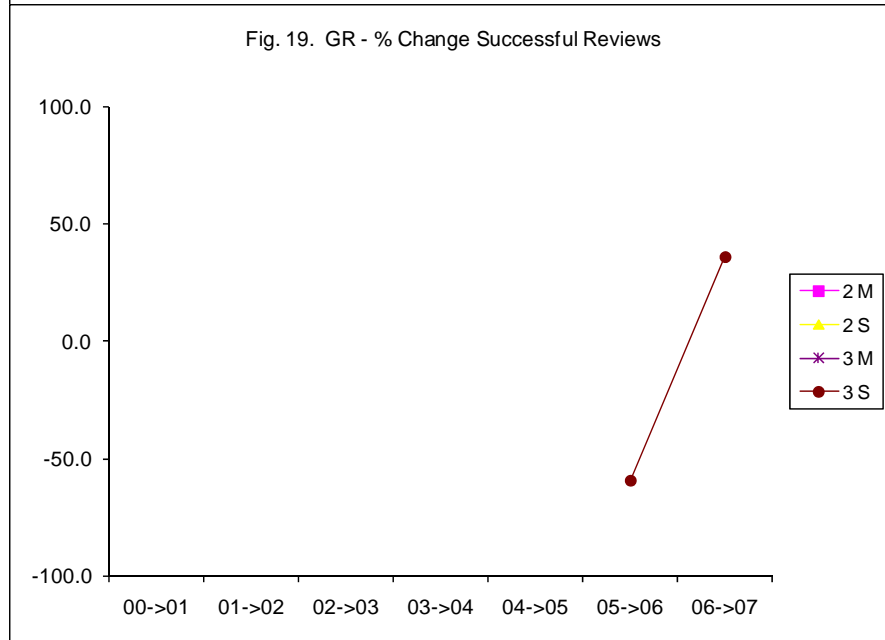
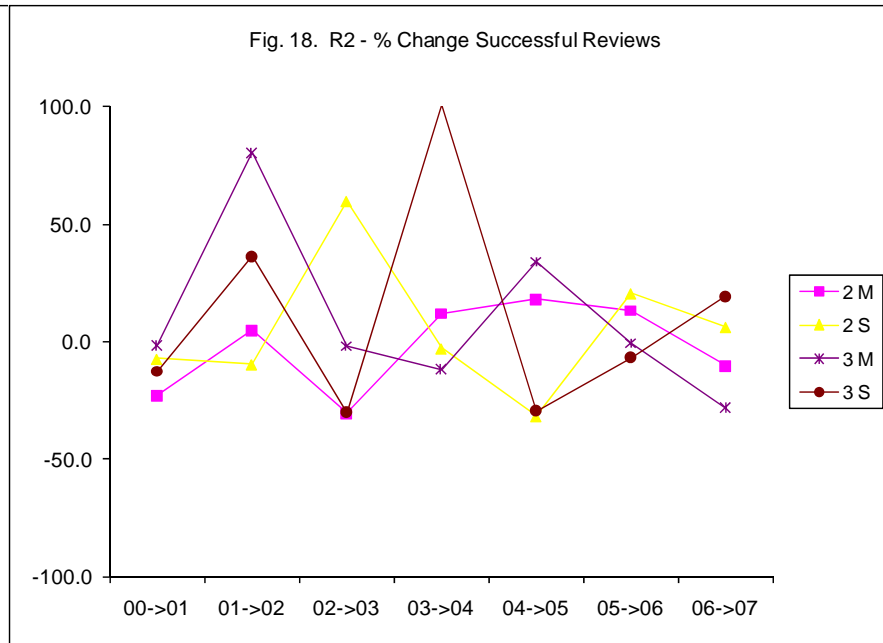
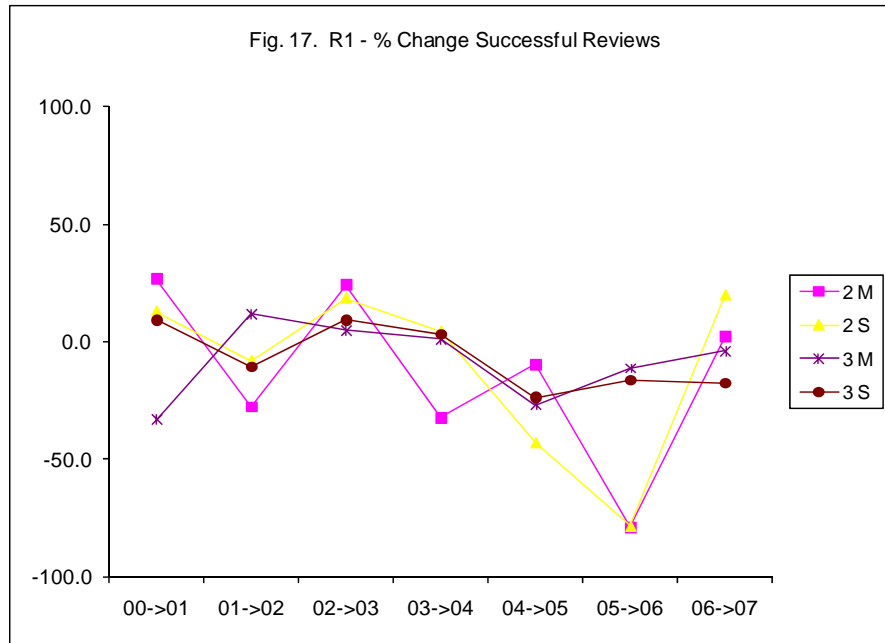
Fig. 4. ALL - No. Review Requests

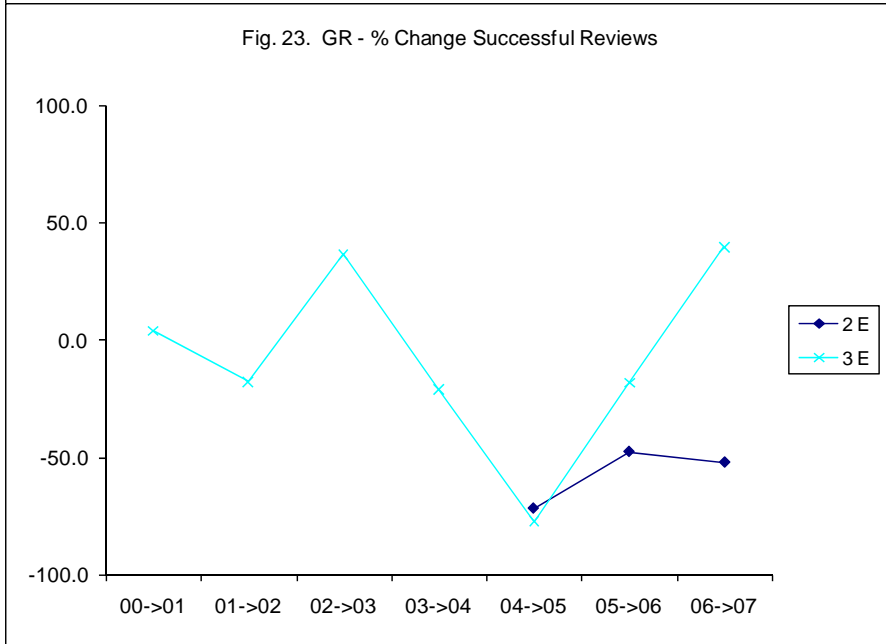
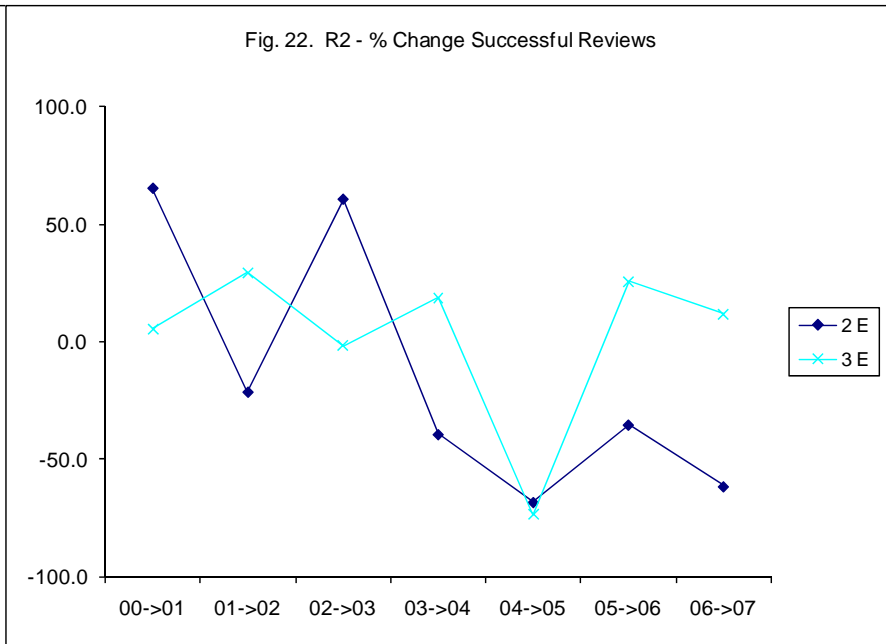
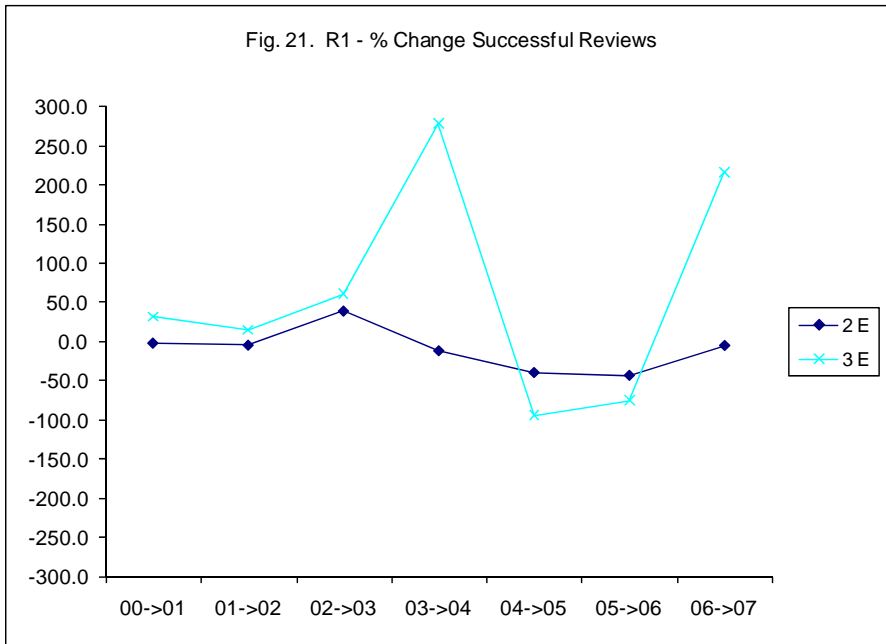


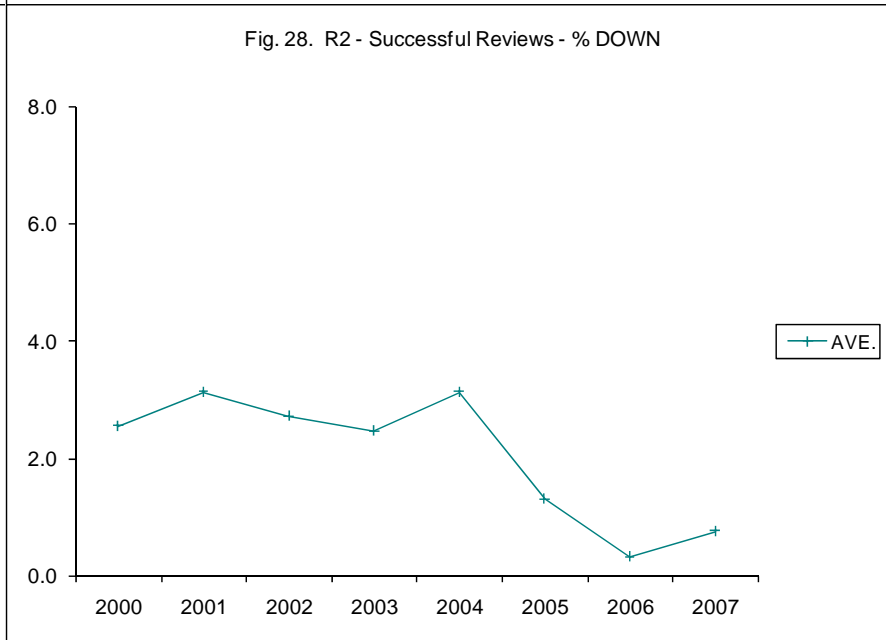
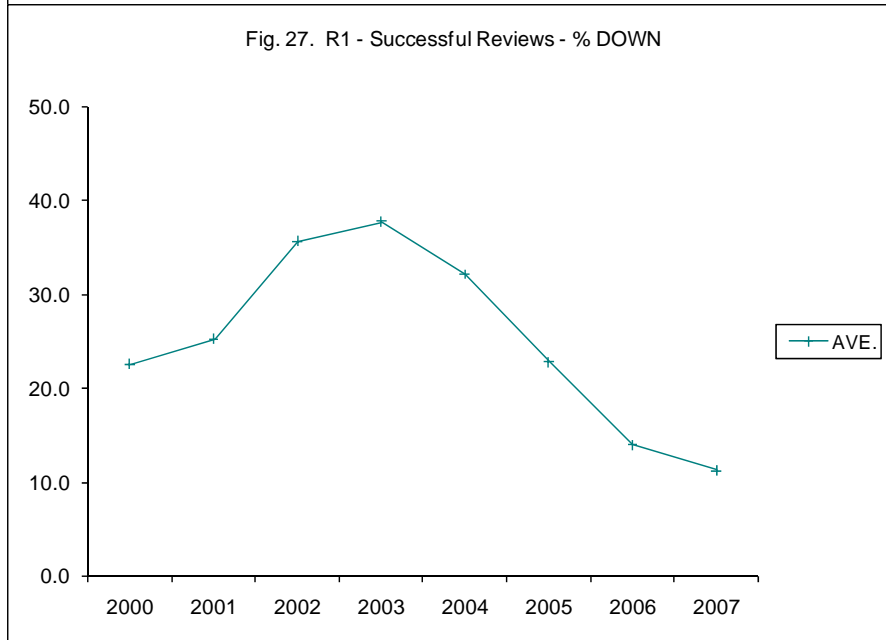
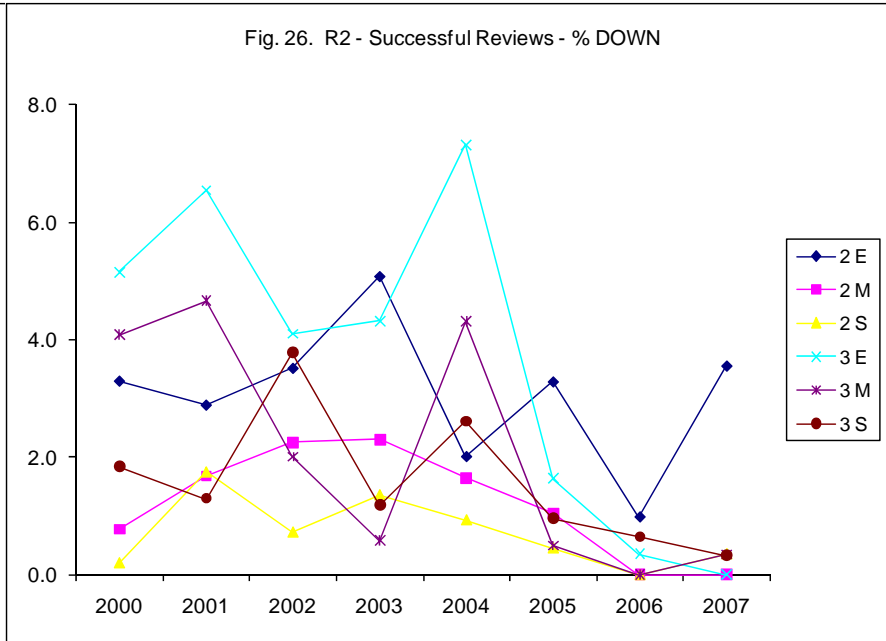
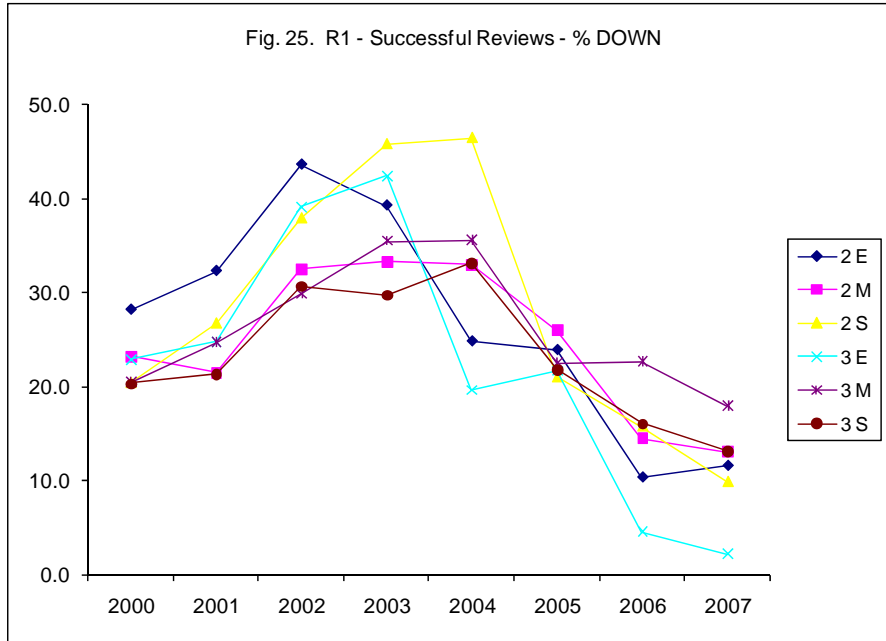


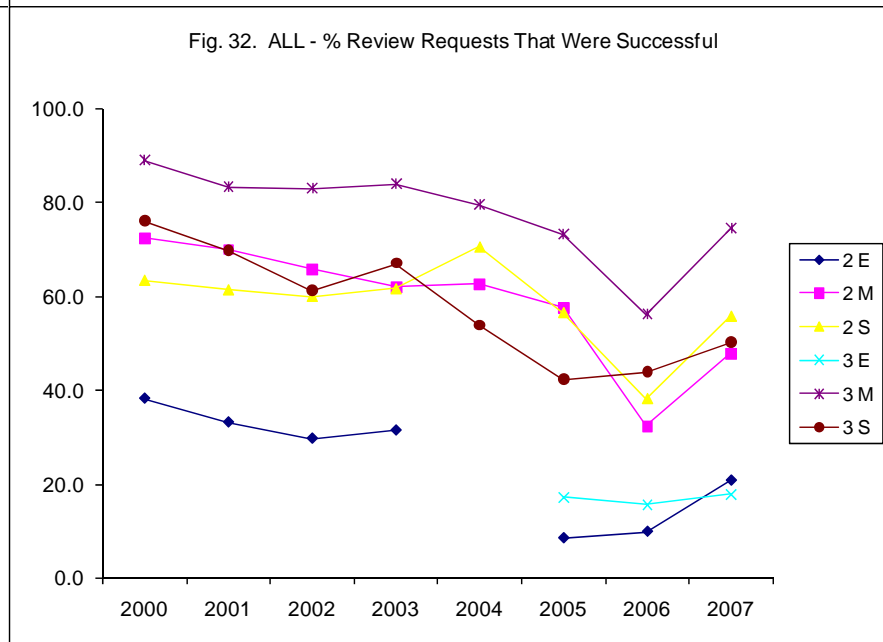
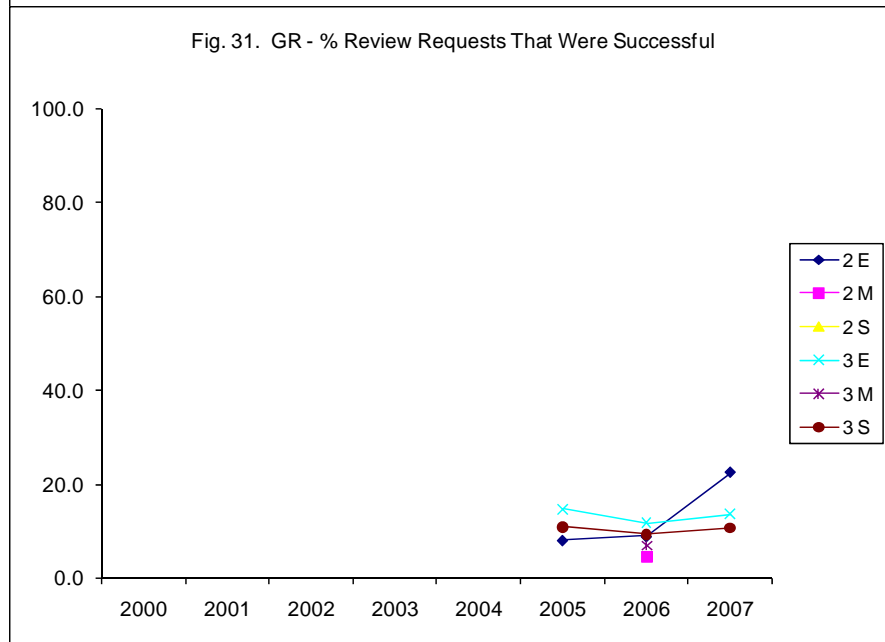
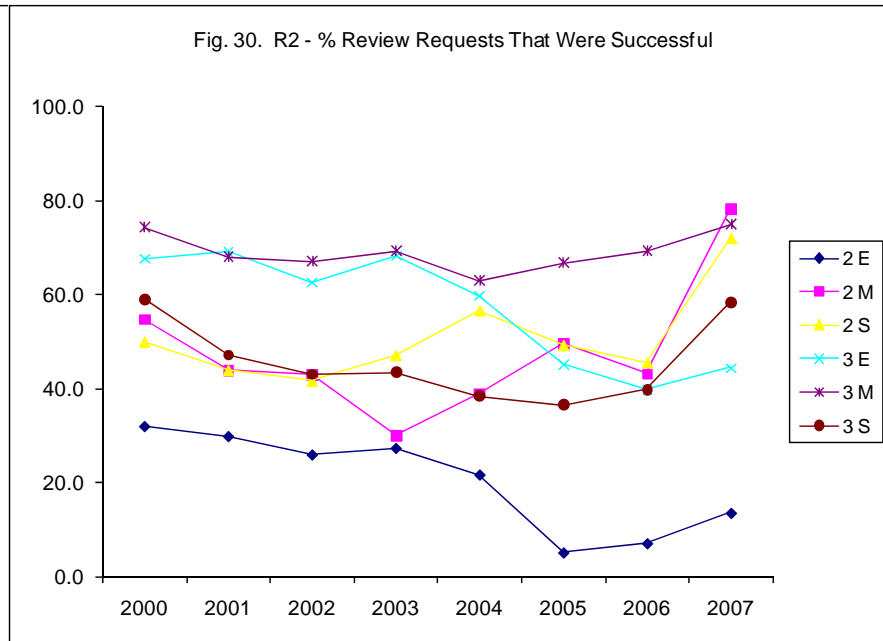
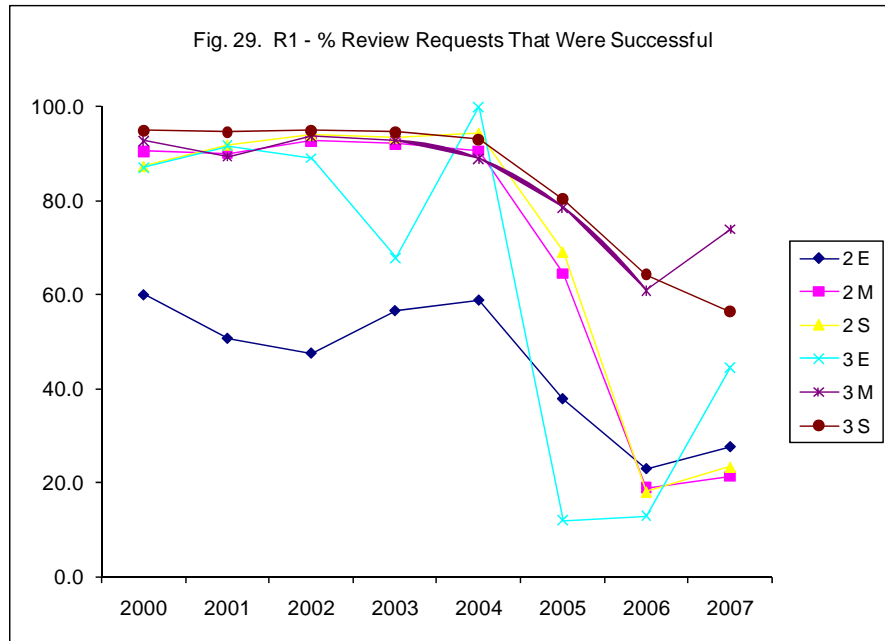












Ofqual wishes to make its publications widely accessible. Please contact us if you have any specific accessibility requirements.

First published by the Office of the Qualifications and Examinations Regulator in 2009.

© Qualifications and Curriculum Authority 2009

Ofqual is part of the Qualifications and Curriculum Authority (QCA). QCA is an exempt charity under Schedule 2 of the Charities Act 1993.

Reproduction, storage or translation, in any form or by any means, of this publication is prohibited without prior written permission of the publisher, unless within the terms of the Copyright Licensing Agency. Excerpts may be reproduced for the purpose of research, private study, criticism or review, or by educational institutions solely for education purposes, without permission, provided full acknowledgement is given.

Office of the Qualifications and Examinations Regulator  
Spring Place  
Coventry Business Park  
Herald Avenue  
Coventry CV5 6UB

Telephone 0300 303 3344  
Textphone 0300 303 3345  
Helpline 0300 303 3346

[www.ofqual.gov.uk](http://www.ofqual.gov.uk)