A NON-ZERO SAMPLING PLAN FOR THE MODERATION OF EXAMINATION PAPERS

W. van Wijck

Department of Industrial Engineering
University of Stellenbosch, South Africa
wvw@sun.ac.za

ABSTRACT

The moderation of examination answer books is an area where quality assurance is essential, and should be employed to ensure that an examination paper's standard, content and span, marking, etc. are fair and reasonable. A scientific procedure is given for finding the minimum number of answer books to moderate (sample size) so that the statement – that no answer book in a set will contain more than a pre-specified proportion of errors – can be made with a pre-specified confidence. The procedure is an extension and enhancement of previous research [7], and guarantees a statistical statement in all cases.

OPSOMMING

Gehalteversekering is belangrik by die moderering van eksamen antwoordstelle om te verseker dat die standaard, inhoud, omvang en akkuraatheid van eksaminering billik en volgens aanvaarbare norme verloop het. 'n Wetenskaplike prosedure word voorgestel waarvolgens die minimum getal antwoordstelle (steekproefgrootte) vir moderering bepaal kan word sodat die stelling dat geen antwoordstel in 'n groep meer as 'n vooraf-gespesifiseerde aantal foute sal bevat nie met 'n vooraf-gespesifiseerde vlak van vertroue gemaak kan word. Die prosedure is 'n uitbreiding en verfyning van vorige navorsing [7], en waarborg in alle gevalle 'n statistiese uitspraak.

1. INTRODUCTION

Moderation within the context of the examination process in an academic institution is a quality assurance activity to establish whether:

- an examination paper has the right standard (level of difficulty)
- an examination paper can be finished within the specified period of time
- an examination paper covers the specified outcomes of a module
- a memorandum for the examination paper exists and whether it is complete, correct, and unambiguous
- individual marks were awarded accurately according to the memorandum (focus of this paper)
- marks were added up correctly
- marks were recorded and processed correctly
- no discrepancies exist between the categories: distinction, pass, re-evaluation, and fail

Each of the above objectives represents a quality characteristic of the examination process. It is therefore possible to do a classification of defects (see next section). To check whether individual marks were awarded accurately according to the memorandum is probably the most difficult, challenging, and time consuming part of the process. It is impractical and uneconomical to expect a moderator to check every individual answer book in the set thoroughly. On the other hand, a moderator should be able to declare a degree of confidence regarding this aspect of the examination process. Clearly the solution is to devise a sampling strategy and procedure that will meet these objectives. This paper proposes a non-zero sampling plan that moderators can use to:

- minimize their amount of inspection;
- declare with a specified confidence whether a specified minimum marking accuracy has been achieved.

Non-zero sampling plans are those where a pre-specified number of defects are allowed in the inspection sample. This article is the final report on research that was done to find a mathematical approach to improve the process of moderation, and represents an extension and enhancement of previous research that was published in this journal [7].

2. CLASSIFICATION OF DEFECTS

In terms of the eight quality objectives (and the corresponding quality characteristics) listed above, examination defects can be classified as shown in Table 1. The table is arranged into four columns, and for each defect-class a preferred inspection strategy is suggested. The various defects are classified into one of two classes: major or minor. No defects are considered critical (this category is normally reserved for life-threatening situations) and no defects are considered so unimportant as to deserve the category of “incidental” defects.

Except for the fifth objective above (checking whether individual marks were awarded accurately according to the memorandum), where non-conformances are classified as minor defects, all other defects are considered to be major. It is ironic that the most time-consuming and difficult objective is the only one with defects that fall into the minor category.

Defect description	Defect class	Suggested inspection method	Reason/ Explanation
• Checking whether an examination paper has the right standard (level of difficulty) • Checking whether an examination paper can be finished within the specified period of time • Checking whether an examination paper covers the specified outcomes of a module • Checking whether a memorandum for the examination paper exists, and whether it is complete, correct, and unambiguous	Major	100% inspection	These types of defects can result in gross errors in the final grading of a student, and will affect all students in the class to a greater or lesser extent
• Checking whether marks were added up correctly • Checking whether marks were recorded and processed correctly • Checking that no discrepancies exist between the categories: distinction, pass, re-evaluation, and fail	Minor		Defects of this type are less important because not all the students in the class are necessarily affected. 100% inspection is still recommended because it is economical for these defects.
• Checking whether individual marks were awarded accurately according to a memorandum (focus of this paper)	Minor	Sampling	Defects of this type are less important because not all the students in the class are necessarily affected. To check for these defects is time-consuming and uneconomical, and therefore sampling inspection is the only practical option.

Table 1: A suggested classification of defects for the examination process

3. PROBLEM FORMULATION

3.1 Statement of objective

The objective is to find a systematic procedure to determine how many answer books (k) from a set of (s) books must be moderated, such that we can be at least (1-β)% certain that none of the answer books will have more than (p')% errors.

3.2 Definition of symbols used in the derivation of the formulae

k	The number of answer books in the sample that was moderated
m	The maximum allowable number of incorrectly-awarded marks in any one answer book (a function of p' and n)
n	The total number of individual marks on the memorandum
p'	The maximum proportion of defects allowed in any one answer book
s	The total number of students in the class (total number of answer books)
β	1 - β is the specified minimum confidence that the moderator must have that the specified maximum proportion of defects was not exceeded in any one of the answer books in the set
γ	A parameter of the uniform distribution that is used as the probability density function to describe the error-probability of a particular department's lecturers

Table 2: Constants

C	The total number of errors in the moderated subset of k answer books, i.e. $C = \sum_{i=1}^{i=k} X_i$
P	The proportion of errors made by an examiner (assumed to be an inherent characteristic of the examiner and the specific set of answer books that he/she examined and constant within and between answer books)
$\{X_i\text{: } i=1..s\}$	The number of incorrectly-awarded marks in the i^{th} answer book in the set of s books. Realizations of X_i are indicated with lower capitals as x_i.
$X_{(k)}$	The maximum number of incorrectly-awarded marks per book found in the moderated sample of size k, i.e. $X_{(k)} = \max\limits_{i} \{X_1, X_2,.., X_i,.., X_k\}$

Table 3: Random variables

4. DERIVATION OF THE THEORY FOR THE NON-ZERO SAMPLING PLAN

Consider an examiner with prior probability distribution $P \sim f_P(p)$, $0 \leq p \leq 1$ of making a mistake ("✓ instead of *X*" or "*X* instead of ✓") for each mark awarded in a paper that counts out of n.[1] In the remainder of this article it will be assumed that P has the following uniform distribution:[2]

$$f_P(p) = 1/\gamma, \qquad \text{where } 0 \leq \gamma \leq 1. \tag{1}$$

We shall further assume that this failure probability is a random variable across the space of all examiners belonging to a group (e.g. an academic department), but constant and fixed for the specific examiner whose set of answer books must be evaluated. More specifically, we shall assume that for a particular examiner probability $P=p$ is fixed for each mark awarded in a specific book, and also constant over all books in the set.

Also assume that s is the total number of answer books in the set available to the moderator. Let X_i be the number of errors in the i^{th} book, where $i=1, ..., s$. We will assume that examiners are fairly consistent and that their errors are independent between different marks in the same book, as well as between different books in the set[3]. Under this assumption $X_i \sim$ binomial(n,P).

Let k be the size of the random sample taken from the above set of s books that must be moderated (this is the parameter that will be optimized). Without loss of generality we will assume that books 1, ...,k were moderated giving realizations of the random variables $X_1 = x_1,\ldots,X_k = x_k$. Since we are considering a non-zero sampling scheme, an upper bound for the number of errors in any one book should be specified. Let this be m. Consequently, if X_i exceeds m in any of the moderated books, the set of books is rejected outright because in this case we will know with certainty that the set does not meet the specification. We now define $C = \sum_{i=1}^{k} X_i$ as the total number of errors found in the sample of k books that were moderated.

[1] It is implicitly assumed that the moderation process itself is error-free, i.e. moderators will neither induce further errors over-and-above those made by the examiners, nor will they miss out on any errors made by the examiners.

[2] The uniform distribution was used for no other reason than its simplicity. During the analysis of the zero sampling plan [7] it was found that the form and parameters of this prior distribution have a relatively small effect on the results. It appeared that the knowledge obtained about the specific examiner during the moderation of the sample and captured by the process of Bayesian statistics contained much more information than the prior assumption about the population to which the examiner belongs. As it turns out (see section 5 of this paper) this is not true for the non-zero sampling plan. The parameter γ has a very pronounced effect on the results. This leads one to suspect also that the form of the prior distribution might not be of negligible importance, and that further experimentation with other distributions is required.

[3] This assumption is questionable, but probably not too unrealistic – especially if an examiner follows the practice of marking one question throughout the set of examination papers and then moving on to the next.

Given these initial definitions, we are now in a position to start our inference on the quality of examination of the remaining *s-k* books.

For a given examiner with *P=p* we can write the conditional probability that he/she made $X_i=x_i$ errors in the i^{th} book as:

$$f_{X_i|P}(x_i \mid p) = P(X_i = x_i \mid P = p) = \binom{n}{x_i} p^{x_i}(1-p)^{n-x_i}, \qquad \text{where } x_i = 0,\ldots,n \tag{2}$$

We now consider the joint conditional probability function that each of the books in the moderated sample has no more than *m* errors, while at the same time the total number of errors in the sample is *c*.

$$f_{X_{(k)},C|P}(m,c \mid p) = P(X_{(k)} \le m, C = c \mid P = p), \qquad \text{where } m = 0,\ldots,n;\ c = 0,\ldots,mk \tag{3}$$

The zero-sampling scheme is the special case when *m*=*c*=0. Table 4 lists all the allowable values of *C* for different values of *m* for the cases where *m* ≤ 3.

		Max. nr. allowable err. (m)				
		0	**1**	**2**	**3**	**...**
C	**0**	✓	✓	✓	✓	...
	...		✓	✓	✓	...
	k		✓	✓	✓	...
	...			✓	✓	...
	2*k*			✓	✓	...
	...				✓	...
	3*k*				✓	...
	...					...

Table 4: Allowable values of *C* (indicated with "✓") for different values of *M*

The evaluation of (3) will now be illustrated for two cases: (i) where *C*=5, *m*=3, $X_{(k)}$≤3, *k*=3, and (ii) where *C*=5, *m*=3, $X_{(k)}$≤3, *k*≥5 respectively. The allowable error combinations for these two cases are shown in Tables 5 and 6.

	Nr. errors per book				**Total (*C*)**
	0	**1**	**2**	**3**	
# books with this # errors	0	1	2	0	5
	0	2	0	1	5
	1	0	1	1	5

Table 5: Allowable error combinations for the case where *m*=3, *c*=5, $X_{(k)}$≤3, *k*=3

For the case of Table 5, we can now evaluate (3) as:

$$
\begin{aligned}
f_{X_{(k)},C|P}\left(3,5\,|\,p\right)=\;&\frac{3!}{1!2!}\left[f_{X_i|P}\left(1\,|\,p\right)\right]\left[f_{X_i|P}\left(2\,|\,p\right)\right]^2+\\
&\frac{3!}{1!2!}\left[f_{X_i|P}\left(1\,|\,p\right)\right]^2\left[f_{X_i|P}\left(3\,|\,p\right)\right]+\\
&\frac{3!}{1!1!1!}\left[f_{X_i|P}\left(0\,|\,p\right)\right]\left[f_{X_i|P}\left(2\,|\,p\right)\right]\left[f_{X_i|P}\left(3\,|\,p\right)\right]
\end{aligned}
\qquad (4)
$$

	Nr. errors per book				**Total**
	0	**1**	**2**	**3**	**(*C*)**
# books with this # errors	*k*-5	5	0	0	5
	k-4	3	1	0	5
	k-3	1	2	0	5
	k-3	2	0	1	5
	k-2	0	1	1	5

Table 6: Allowable error combinations for the case where *m*=3, *C*=5, $X_{(k)}$≤3, *k*≥5

For the case of Table 6, we can now evaluate (3) as:

$$
\begin{aligned}
f_{X_{(k)},C|P}\left(3,5\,|\,p\right)=\;&\frac{k!}{5!(k-5)!}\left[f_{X_i|P}\left(0\,|\,p\right)\right]^{k-5}\left[f_{X_i|P}\left(1\,|\,p\right)\right]^5+\\
&\frac{k!}{1!3!(k-4)!}\left[f_{X_i|P}\left(0\,|\,p\right)\right]^{k-4}\left[f_{X_i|P}\left(1\,|\,p\right)\right]^3\left[f_{X_i|P}\left(2\,|\,p\right)\right]+\\
&\frac{k!}{1!2!(k-3)!}\left[f_{X_i|P}\left(0\,|\,p\right)\right]^{k-3}\left[f_{X_i|P}\left(1\,|\,p\right)\right]\left[f_{X_i|P}\left(2\,|\,p\right)\right]^2+\\
&\frac{k!}{1!2!(k-3)!}\left[f_{X_i|P}\left(0\,|\,p\right)\right]^{k-3}\left[f_{X_i|P}\left(1\,|\,p\right)\right]^2\left[f_{X_i|P}\left(3\,|\,p\right)\right]+\\
&\frac{k!}{1!1!(k-2)!}\left[f_{X_i|P}\left(0\,|\,p\right)\right]^{k-2}\left[f_{X_i|P}\left(2\,|\,p\right)\right]\left[f_{X_i|P}\left(3\,|\,p\right)\right]
\end{aligned}
\qquad (5)
$$

Upon expansion of these formulas it becomes clear that for any *m* and *c* we have a function of the form:

$$f_{X_{(k)},C|P}\left(m,c\,|\,p\right)=Ap^c\left(1-p\right)^{nk-c}, \qquad (6)$$

where *A=A(m,c,n,k)* itself is a function of *m*, *c*, *n* and *k* but not *p*.

If we now take the sum over *C* in equation 6 we obtain a density function, conditional on *p*, describing the probability that each of the books in the moderated sample will have no more than *m* errors.

$$f_{X_{(k)}|P}\left(m\,|\,p\right)=\sum\nolimits_{c=0}^{mk}Ap^c\left(1-p\right)^{nk-c} \qquad (7)$$

We now wish to find the posterior distribution of *P*, i.e. $f_{P|X_{(k)}}(p\,|\,m)$, but we first need to determine the probability distribution of $X_{(k)}$:

$$\begin{aligned} f_{X_{(k)}}(m) &= \int_0^{\gamma} f_P(p).f_{X_{(k)}|P}(m\,|\,p).dp \\ &= (1/\gamma)\int_0^{\gamma}\sum_{c=0}^{c=mk} Ap^c(1-p)^{nk-c}.dp \end{aligned} \tag{8}$$

Before proceeding with our discussion it is important to note that (8) can be used to calculate the probability that all books within the moderated sample will conform to the quality criterion (have errors less than or equal to *m*). This has been calculated for the above example, with n =100, γ =0,01 and m=3; and the following results were obtained:

98.76% for the case *k*=3
96.64% for the case *k*=5

We now proceed with the derivation of the posterior distribution of *P* by applying Bayes' formula as follows:

$$\begin{aligned} f_{P|X_{(k)}}(p\,|\,m) &= \frac{f_{X_{(k)},P}(m,p)}{f_{X_{(k)}}(m)} = \frac{f_P(p)\,f_{X_{(k)}|P}(m\,|\,p)}{f_{X_{(k)}}(m)} \\ &= \frac{(1/\gamma)\sum_{c=0}^{c=mk} Ap^c(1-p)^{nk-c}}{(1/\gamma)\int_0^{\gamma}\sum_{c=0}^{c=mk} Ap^c(1-p)^{nk-c}\,dp} \\ &= \frac{\sum_{c=0}^{c=mk} Ap^c(1-p)^{nk-c}}{\int_0^{\gamma}\sum_{c=0}^{c=mk} Ap^c(1-p)^{nk-c}\,dp} \end{aligned} \tag{9}$$

To avoid confusion we will use the symbol Π in reference to the posterior distribution of *P*. Therefore (9) becomes,

$$f_{\Pi}(\pi) = \frac{\sum_{c=0}^{c=mk} A\pi^c(1-\pi)^{nk-c}}{\int_0^{\gamma}\sum_{c=0}^{c=mk} Ap^c(1-p)^{nk-c}\,dp} \quad 0 \le \pi \le \gamma \tag{10}$$

It is easy to show that if *k*=0 in the above equation (i.e. no moderation took place) then the posterior distribution reverts back to the prior (uniform) distribution. This means that no additional knowledge has been acquired about the quality of the specific examiner, and one is only left with the original presumptions about the quality of the department's teaching staff (which of course must be the case).

We now turn our attention to the remaining *s-k* books. From moderating the sample of *k* books, we have gained knowledge about the accuracy of the specific examiner. With this knowledge, the probability that this examiner made *x* mistakes in any one of remaining *s-k* books can now be stated as:

$$f_{X_i|\Pi}(x\,|\,\pi)=P(X_i=x\,|\,\Pi=\pi)=\binom{n}{x}\pi^x(1-\pi)^{n-x} \qquad \text{where} \quad x=0,\ldots,n \quad \text{and} \quad i=k+1,\ldots,s \tag{11}$$

The probability that the proportion of defects in any one of the remaining *s-k* books will be equal to *x/n* is therefore also given by (11) above. The chance that this proportion will be less than a pre-specified proportion *p′* can therefore be found as follows:

$$\begin{aligned} F_{X_i|\Pi}(m\,|\,\pi) &= \sum\nolimits_{x=0}^{m}\binom{n}{x}\pi^x(1-\pi)^{n-x} \quad \text{where } m=\lfloor np'\rfloor \le np' \\ &= (1-\pi)^n\sum\nolimits_{x=0}^{m}\binom{n}{x}\left(\frac{\pi}{1-\pi}\right)^x \end{aligned} \tag{12}$$

Given our assumption of independence, we can extrapolate (12) to find the probability that all the remaining books meet the quality criterion *p′*:

$$\left[F_{X|\Pi}(m\,|\,\pi)\right]^{s-k} \text{ where } m=\lfloor np'\rfloor \le np' \tag{13}$$

Since the expression in (13) is a random variable with respect to Π we can take its expected value to find the expected probability (over the population of all examiners) that all the remaining books meet the quality criterion *p′*.

$$\begin{aligned} E\left[F_{X|\Pi}(m\,|\,\pi)\right]^{s-k} &= \int_0^{\gamma}\left[\frac{\sum_{c=0}^{c=mk}A\pi^c(1-\pi)^{nk-c}}{\int_0^{\gamma}\sum_{c=0}^{c=mk}Ap^c(1-p)^{nk-c}\,dp}\right]\left[F_{X|\Pi}(m\,|\,\pi)\right]^{s-k}d\pi \\ &= \frac{\int_0^{\gamma}\sum_{c=0}^{c=mk}A\pi^c(1-\pi)^{nk-c}\left[F_{X|\Pi}(m\,|\,\pi)\right]^{s-k}d\pi}{\int_0^{\gamma}x\sum_{c=0}^{c=mk}Ap^c(1-p)^{nk-c}\,dp} \\ &= \frac{\int_0^{\gamma}\sum_{c=0}^{c=mk}A\pi^c(1-\pi)^{ns-c}\left[\sum_{x=0}^{m}\binom{n}{x}\left(\frac{\pi}{1-\pi}\right)^x\right]^{s-k}d\pi}{\int_0^{\gamma}\sum_{c=0}^{c=mk}Ap^c(1-p)^{nk-c}\,dp} \\ &= \frac{\int_0^{\gamma}(1-\pi)^{ns}\left[\sum_{x=0}^{m}\binom{n}{x}\left(\frac{\pi}{1-\pi}\right)^x\right]^{s-k}\left[\sum_{c=0}^{c=mk}A\left(\frac{\pi}{1-\pi}\right)^c\right]d\pi}{\int_0^{\gamma}\sum_{c=0}^{c=mk}Ap^c(1-p)^{nk-c}\,dp} \ge 1-\beta \end{aligned} \tag{14}$$

In the above inequality β is the chance that the criterion will not be met in at least one of the remaining books even though the moderator encountered no book in the sample of k books with more than m errors. The value of 1 - β can therefore be regarded as the confidence that the prescribed accuracy was achieved by the examiner. The left hand side of the inequality therefore represents the confidence in the quality of examination acquired through moderation, while the right hand side represents the minimum required confidence (i.e. the standard). If C is forced not to exceed zero (C=0) the sampling plan defaults to the zero sampling plan discussed in [7], and it can be shown that the above inequality reduces to equation (10) in [7]. The reader will also notice that if the entire set of books is moderated (k=s), then the acquired confidence is 100% (which of course must be the case). The inequality can be solved for k using numerical integration. The smallest k that satisfies the inequality is recommended as the sample size for moderation.

5. CHARACTERISATION OF THE SAMPLING PLAN AND ITS PARAMETERS

A MatLab program was written to solve inequality (14) for a range of the input parameter values that covers a wide spread of real life scenarios.[4] The results are tabulated in Tables 7 and 8. The following ranges of values were used for the respective input parameters:

p': 0,00 to 0,05 (no fixed increments; increments are dictated by the choice of m-values). The lower bound corresponds with a standard that allows no errors in any of the answer books (very strict), while the upper bound corresponds with a standard that allows up to 5% errors in any individual answer book (very "loose"). When it is customary during the final grading process to round up by 2,5% (e.g. an achieved mark of 47,5% may be rounded up to a final mark of 50%), the value of this parameter must be small enough not to severely affect the outcome of this practice. Suggested values for this parameter are between 0,02 and 0,03.

s: 20 to 100 in increments of 10. The lower bound corresponds with a class size of 20 while the upper bound corresponds with a class size of 100. University class sizes are seldom smaller than 20 but there are many that exceed 100. The run time of the algorithm becomes long for large class sizes, and this is why an upper bound of 100 was chosen for this paper.

n: 20 to 100 in increments of 20. The lower bound corresponds with a memorandum having 20 marks, while the upper bound corresponds with one that has 100 marks. It is believed that this range covers most of the scenarios encountered in practice.

β: 0,15. This relatively large (single) value for the required confidence in the quality of the examination process was chosen to obtain a satisfactory trade-off between (i) the amount of moderation that is required, and (ii) the amount of confidence that is needed. Smaller values for this parameter (e.g. 0,05 or 0,10) result in amounts of moderation that are clearly uneconomical.

[4] I would like to use this opportunity to thank my son Tjaart for the many hours he devoted to writing the MatLab program and conducting the very time-consuming computer runs.

γ: 0,01 and 0,02. The data in Table 7 are for γ=0,01 while that in Table 8 are for γ=0,02. $\gamma/2$ represents the average proportion of errors an arbitrary member of the teaching staff of department is expected to make. γ=0,01 therefore refers to a department where the average proportion of errors of the teaching staff is in the region of 0,005 (0,5%). This means that the "average" lecturer will only make one mistake in every two answer books with a memorandum that contains 100 marks. γ=0,02 allows for twice this amount. Nonetheless these are very small values which require very accurate marking by the teaching staff. The results indicate that the amount of moderation necessary to obtain the required confidence is very sensitive to the accuracy of a department's teaching staff. Higher values of γ will therefore result either in relatively low confidence (large β) or in high volumes of moderation (large k), or both.

A glimpse at the results of Tables 7 and 8 reveals many interesting and often complicated relationships between the different input parameters. It is not our intention to discuss all these relationships here; instead the interested reader is encouraged to study the Tables in more detail. However, the following important general relationships and conclusions deserve mention:

1) The required number of papers that must be moderated (k) is almost linearly proportional to the class size (s). Bigger classes require more moderation.

2) The strictness of the quality standard for examination (p') has a very strong impact on the required amount of moderation (k). For example, if no errors are allowed in any examination answer book, then for a class size of 100, a memorandum with 100 marks, and a group of teachers for which γ=0,02, 86 out of the 100 papers must be moderated to obtain a confidence of more than 85% in the quality of the marking process. If 5% errors are allowed with the remaining input parameters unchanged, then only 5 examination books need to be moderated to obtain the same degree of confidence. The reader will also notice that the *number* of errors allowed per answer book (m), rather than the *proportion* of errors (p'), is the dominant factor for the amount of moderation that needs to take place (k). In practice, p' is likely to be set below 2,5%. The reader will notice that smaller memorandums (small n) require more moderation (larger k) than larger ones. This is because more is learned per book about the quality of the examiner in the case of a memorandum containing many marks (large n). Bayes learning is steeper in this case than for memorandums with fewer marks. It is clear that the total number of individual marks moderated (i.e. nk) strongly determines the amount of "learning" that takes place during moderation.

3) Lastly, the quality of the teaching staff, represented by the parameter γ, strongly affects the amount of moderation that needs to be done. It is alarming to see how accurate a department's teaching staff needs to be to get anywhere near the standards we have taken for granted. This is probably the most revealing conclusion that came out of this study!

A 3-dimentional plot of the results of Table 7 is shown in Figure 1. Although this figure is not particularly useful as a source for reading off k-values, it does graphically illustrate the general relationship between the various input parameters (γ excluded).

			s								
N	*m*	*p′*	20	30	40	50	60	70	80	90	100
20	0	0.00%	17	26	34	43	52	60	69	77	86
	1	5.00%	1	1	9	19	27	35	44	52	61
40	0	0.00%	17	26	35	43	52	60	69	77	86
	1	2.50%	11	19	28	38	43	51	59	66	74
	2	5.00%	0	0	0	0	0	0	0	0	8
60	0	0.00%	18	26	35	43	52	60	69	77	86
	1	1.67%	14	22	30	37	45	52	59	67	74
	2	3.33%	1	1	9	17	25	33	40	48	55
	3	5.00%	0	0	0	0	0	0	0	0	0
80	0	0.00%	18	26	35	43	52	60	69	77	86
	1	1.25%	15	23	30	37	45	52	59	67	74
	2	2.50%	6	14	21	29	36	43	50	57	64
	3	3.75%	0	0	0	0	0	0	0	0	3
	4	5.00%	0	0	0	0	0	0	0	0	0
100	0	0.00%	18	26	35	43	52	60	69	77	86
	1	1.00%	15	23	30	37	45	52	59	67	74
	2	2.00%	10	18	25	32	39	46	52	59	65
	3	3.00%	0	0	0	5	7	9	10	10	11
	4	4.00%	0	0	0	0	0	0	0	0	0
	5	5.00%	0	0	0	0	0	0	0	0	0

Table 7: k-values for β=0,15 and γ=0,01

			s								
n	*m*	*p′*	20	30	40	50	60	70	80	90	100
20	0	0.00%	17	26	35	43	52	60	69	77	86
	1	5.00%	11	19	27	35	43	51	59	66	74
40	0	0.00%	18	26	35	43	52	60	69	77	86
	1	2.50%	15	23	30	37	45	52	59	67	74
	2	5.00%	5	13	21	29	36	43	50	56	62
60	0	0.00%	18	26	35	43	52	60	69	77	86
	1	1.67%	15	23	30	37	45	52	61	67	74
	2	3.33%	12	20	26	33	40	46	53	59	65
	3	5.00%	1	4	7	8	9	10	11	11	12
80	0	0.00%	18	26	35	43	52	60	69	77	86
	1	1.25%	15	23	30	37	45	52	59	67	74
	2	2.50%	14	20	27	33	40	46	53	59	65
	3	3.75%	6	8	9	10	10	11	11	12	12
	4	5.00%	0	0	2	4	5	6	7	7	7
100	0	0.00%	18	26	35	43	52	60	69	77	86
	1	1.00%	15	23	30	38	45	52	59	67	74
	2	2.00%	14	20	27	33	40	46	53	59	65
	3	3.00%	7	8	9	10	10	11	11	12	12
	4	4.00%	2	5	6	6	7	7	7	8	8
	5	5.00%	0	0	0	0	0	0	4	4	5

Table 8: k-values for β=0,15 and γ=0,02

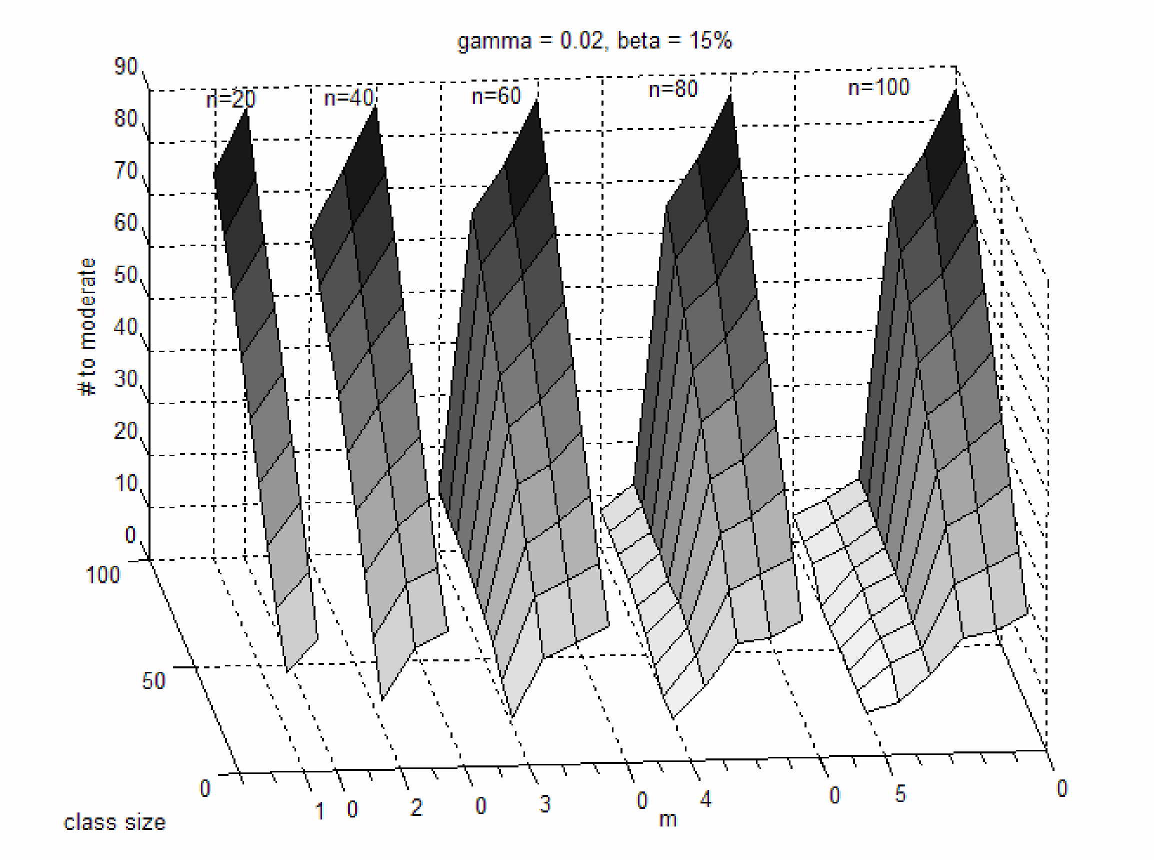

Figure 1: A plot of the data of Table 7

6. CONCLUSION

As elsewhere, educational institutions in South Africa experience increasing pressure from stakeholders and regulatory bodies to follow procedures that will ensure a quality service. This article is the final report on research that was conducted at the Department of Industrial Engineering of the University of Stellenbosch to put the process of moderation on a more scientific footing.

The proposed method allows a moderator to choose a realistic and practical number of answer books from a set of books that must be evaluated. Then, based on the total number of errors that the moderator found in the sample, a statement whether or not the quality of the examiner's marking meets the minimum requirement can be made with a specified and known level of confidence. The following five input parameters are accounted for:

n: The number of marks on the memorandum
p': The maximum proportion of errors allowed in any one answer book
s: The number of answer books in the set (class size)
β: The required confidence (1-β) in the quality of the examination process
γ: The presumed accuracy of the population of teaching staff

Because the relationship between the various parameters is complex, a MATLAB computer program was written that moderators can now use as a tool. Some of the underlying assumptions that were used in the derivation of the sampling plan are indeed untested and questionable, as was pointed out at appropriate places in the text. However, the theory provides us with a ballpark sample size based on scientific reasoning – and herein, perhaps, lies its greatest value.

Although the research was specifically driven by the desire to improve the quality of a very important educational process (assessment), the research result is a sophisticated sampling scheme that may very well have wider application, particularly in the engineering field.

7. REFERENCES

[1] **Kuo, T. and Mital, A.** 1993. Quality control expert systems: A review of pertinent literature. *Journal of Intelligent Manufacturing Systems*, 4: pp. 245-257.

[2] **Mital, A., Nicholson, A.S. and Ayoub, M.M.** 1993. *A guide to manual materials handling*. Taylor & Francis, Ltd, London, United Kingdom.

[3] **Mital, A. and Anand, S** 1993. Insignia: Insignia Solutions home page. *Handbook of expert systems in manufacturing: Structure and rules*. Chapman & Hall, London, United Kingdom.

[4] Java Home Page. http://java.sun.com/

[5] **Mital, A.** 1988. *Desirability of robots.* In *International Encyclopedia of Robotics* (ed. R.C. Dorf). Wiley-Interscience, New York, pp. 322-329.

[6] **Mital, A. and Mahajan, A.** 1989. *Impact of production volume and wage and interest rates on economic decision making: The case of automated assembly. Proceedings of the Conference of Society for Integrated Manufacturing*, Institute of Industrial Engineers, pp. 558-563.

[7] **Van Wijck, W. and Dirkse van Schalkwyk, T.** 2005. A zero sampling plan for the moderation of examination papers, *South African Journal of Industrial Engineering*, 16(2), pp 69-80.