NASA CONTRACTOR REPORT

NASA CR-1784

MANIPULATION ERRORS IN COMPUTER SOLUTION OF CRITICAL SIZE STRUCTURAL EQUATIONS

by R. J. Melosh

Prepared by
PHILCO-FORD CORPORATION
Palo Alto, Calif.
for Goddard Space Flight Center

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION • WASHINGTON, D. C. • JUNE 1971

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
NASA CR-1784		

4. Title and Subtitle	5. Report Date
MANIPULATION ERRORS IN COMPUTER SOLUTION OF CRITICAL SIZE STRUCTURAL EQUATIONS	June 1971
	6. Performing Organization Code
7. Author(s)	8. Performing Organization Report No.
R. J. Melosh	
	10. Work Unit No.
9. Performing Organization Name and Address	
Philco-Ford Corporation Palo Alto, California	11. Contract or Grant No. NAS5-10365
	13. Type of Report and Period Covered
12. Sponsoring Agency Name and Address	Contractor Report
National Aeronautics and Space Administration Washington, D.C. 20546	14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

Errors that arise from digital computer solutions of structural problems using the finite element method are explored. It reports and interprets measured errors for problems involving the number of calculations ranging up to the critical number, i.e. the base 2 raised to the power equal to the bits in the computer word. Structures whose connections are in series, parallel, and a combination of both are examined against eight error measures. Error measures are found to be sensitive to error measuring and problem details such as loading, sequencing, and the definition of norm. Four error measures are recommended for the displacement method analysis.

17. Key Words (Suggested by Author(s))	18. Distribution Statement
Numerical Methods, Error Analysis, Finite Element Analysis, Structural Analysis, Matrix Conditioning, Computer Errors	Unclassified - Unlimited

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price*
Unclassified	Unclassified	48	$3.00

*For sale by the National Technical Information Service, Springfield, Virginia 22151

FOREWORD

This report is prepared under Contract NAS5-10369. The NASA program monitor is Mr. Thomas G. Butler. Research for this report was accomplished in the period December 1969 through November 1970 and the report was submitted in December 1970.

MANIPULATION ERRORS IN COMPUTER SOLUTION OF CRITICAL SIZE STRUCTURAL EQUATIONS

by R. J. Melosh*

SUMMARY

The document supplements previous studies of computer-induced manipulation errors in finite element analyses of structures. It reports and interprets measured errors for problems ranging in size up to the critical number of calculations, 2^p where p is the computer precision. p is 27 for these problems.

It cites errors measured in analyzing mixed structural systems composed of beam elements. It defines eight error measures and evaluates their behavior as a function of the definition of the norm, number of calculations, matrix density, number of equations, use of critical number components, and minor arithmetic sequence changes. The measures provide for distinguishing the relative value of residual and solution measures and the relative importance of error sources.

It concludes that the error bounds based on the number of calculations is an increasingly poor guide to actual errors as the number increases. It recommends development of statistical data relating the number of calculations and error from practical problems. This data would be used to indicate the probability of success for particular problems.

It indicates that error measures are sensitive to error measuring and problem details. Measures of residual errors are much greater than corresponding solution error measures. Change of loading, minor changes in the arithmetic sequence, and change in the definition of the norm--each has a significant effect on indicated errors.

It reports that inherited and decomposition errors are the more important error sources. It compares discretization and manipulation errors and finds discretization more important in small problems (less than 2000 equations on a 27 bit computer).

It recommends four error measures to be incorporated in displacement method analyses; a numerical singularity check, positive definite checks, a total solution energy error measure, and a stress precision check. These are selected because of their demonstrated validity, economy, and ability to distinguish among important error sources.

TABLE OF CONTENTS

FIGURES AND TABLES

Figures

Tables

Section 1

INTRODUCTION

This is the third in a series of reports on manipulation errors induced by the digital computer in finite element analyses of structures. The first report[1]* surveys all the errors induced in the computer phase of analyses. The second report[2] examines the effect on error magnitudes of changing various analysis parameters. This pair of reports includes data for numerical experiments on a 27 bit mantissa computer with up to 1200 equations and less than the critical number of calculations: 13.4×10^6.

This report describes the results of additional experiments for sets of up to 2000 equations and 40×10^6 calculations. These data pertain to frame structures with varying degrees of connectivity. One group of tests yields data on relative error magnitudes in parts of the solution process and comparisons of error criteria. The second group shows the effect of the number calculations and calculation sequence changes when the critical number of calculations are involved.

This report describes the basis of experimentation, the results and their interpretation. The next section cites the test process and error criteria and justifies the experimental setup common to all testing. The third section describes the experiments for the problems involving less than the critical number of calculations. The fourth section cites data relevant to the tests for critical problem size analysis. The fifth section identifies error checks recommended for production computer analyses. The last section contains conclusions.

The assistance of Robert Steeley of Philco-Ford in modifying existing computer codes, and managing experimental data development was vital in this study. In addition, the assistance of the Ames Research Center Computer Laboratory in implementing experiments, is gratefully acknowledged.

*Superscripts in parentheses denote numbers of references cited in REFERENCES.

Section 2

BASIS OF EXPERIMENTS

This section describes the test problems and testing procedures for extending study of manipulation errors into the critical problem range. It details error measures used. It justifies the experimental basis using results and conclusions from Ref. 1 and 2.

Analysis Approach

The analysis method for selected problems will use finite elements and the direct stiffness method. In this method, stiffness coefficients and stress coefficients are developed directly to express the load-deflection and stress displacement equations in the form:

$$K\Delta = F \tag{2-1}$$

$$\sigma_i = S_i \Delta \qquad i = 1, 2, \ldots f \tag{2-2}$$

where K is the symmetric, positive definite stiffness matrix,
Δ is the vector of displacement components,
F is the vector of joint loads,
σ is the vector of stresses for element i,
S_i is the matrix of stress coefficients for element i, and
f is the number of finite elements in the structural system.

Equations (2-1) are solved for displacements. The solution will be developed by decomposing K into the form:

$$L D^{-1} L^T = K \tag{2-3}$$

where L is a lower triangular matrix,
D^{-1} is a diagonal matrix, and
L^T is the transpose of L.
Forward substitutions will evaluate y where

$$Ly = F \tag{2-4}$$

Division by the diagonals will then form z where

$$z = Dy \tag{2-5}$$

Backward substitutions will produce Δ by solving

$$L^T \Delta = z \tag{2-6}$$

With the displacements known, the stresses would be found by the multiplication of Eq. (2-2). Since stresses are differentials of displacements, this multiplication is intrinsically a differencing operation.

It has been shown in Ref. 1, page 135-137, that the largest error in both the displacement and force methods, in solving for primary unknowns, involves a positive definite matrix of coefficients. In the displacement method, this is the stiffness matrix. In the force method, this is the redundants' matrix. This matrix remains after displacement unknowns are eliminated from the structural equations.

Furthermore, Ref. 2, page 103, reports that the same error sources limit analysis accuracy in both methods. Large singularity errors are possible in each and are avoidable by care in calculation sequencing. In the displacement method this is achieved by taking care in sequencing joints. In the force method, this is achieved by careful selection of force redundants. Critical arithmetic errors can arise in each method. In the displacement method these can occur in developing stresses by differencing displacements. In the force method, these can arise in adjusting the value of internal forces in the elements of the determinate substructure due to the redundants.

Thus, examination of errors in the displacement method will study the same type of equations as cause manipulation error difficulties in the force method. The principal error sources in stiffness equation solution have their counterpart in the force method.

The modified Gauss decomposition defined by Eqs. (2-3) is selected to avoid errors introduced by taking square roots, as required in Choleski decomposition. These errors can be much larger than the error of one part in the last binary position possible in the diagonal divisions. Reference 1, page 41-47, shows these errors can result in numerical instability. Reference 3, page 667, reports a test of 200 joint regular cantilevered beam in which these errors result in 38% error in displacement predictions using a 27 bit mantissa computer.

Test Problems

All tests use the straight prismatic beam stiffness matrix as the basic and only element stiffness matrix. This matrix relates a set of forces at the ends to corresponding displacements by the equation:

$$\begin{Bmatrix} F_{zi} \\ \frac{1}{a}M_{yi} \\ F_{zj} \\ \frac{1}{a}M_{yj} \end{Bmatrix} = \frac{2EI}{a^3}\begin{bmatrix} 6 & & & \\ 3 & 2 & & \\ -6 & -3 & 6 & \\ 3 & 1 & -3 & 2 \end{bmatrix}\begin{Bmatrix} u_{zi} \\ a\theta_{yi} \\ u_{zj} \\ a\theta_{yj} \end{Bmatrix} \qquad (2\text{-}7)$$

where F is a force,
M is a moment,
a is the beam length,
E is Young's Modulus,
I is the bending moment of inertia,
u is a lateral displacement,
θ is a rotation in bending,
The first subscript denotes the vector direction (the axis coincides with the beam neutral axis), and the second subscript denotes the beam end.

All tests represent the stiffness coefficients by the nondimensional coefficients given in Eq. (2-1). All segments are required to be the same length. The total stiffness matrix is therefore formed simply as the union of element stiffness matrices. When stiffness variations are considered, the scalar stiffness, $\frac{2EI}{a^3}$, is changed and the nondimensional coefficients scaled. These scalars are entered as integers so that only the relative magnitude of the stiffness scalars affects the manipulation errors.

Structural systems modelled consist of collinear beam segments. Each joint is connected to one or more sequentially higher numbered joints as required to provide stiffness matrices of various densitites. The sparsest of these matrices represents a cantilevered beam. A completely full matrix represents a system in which every joint is connected to every other joint. This class of matrices is called a "mixed structural system" because elements are neither all acting in series nor all in parallel. Since the series system is the sparsest matrix and the all parallel is full, the mixture can be represented by the percent of non-zero elements in the matrix, the matrix population density.

Table I particularizes the form of the stiffness matrices for all mixed structural systems. In test problems, the K_{11}, K_{12}, and K_{22} partitions are those indicated as partitions in Eq. (2-7). The total unrestrained stiffness matrix is characterized by the matrix order, its bandwidth, and the geometric progression ratio. If this ratio is 1.0, the structure is "regular". For this class of matrices, the solution wavefront and matrix bandwidth are the same.

The beam element stiffness matrix is chosen as the basic unit because previous study shows it involves larger manipulation errors than rods, membranes, or prisms. Decomposition errors for cantilevered beams were shown to vary as $f^4 b^{-p}$ where f is the number of finite elements; b the number base; and p the computer precision. Decomposition error for a straight hinged rod, on the other hand, varies as fb^{-p}. In the worst case, singularity error limits the number of beams to $fb^{-(p-2)/3}$ but fb^{-p} rod sigments can be treated. Substitution errors vary as $360fb^{-p}$ for cantilevered beams, but only as $265\,fb^{-p}$ for rods. (p-27). The force equilibrium equations are written first in Eq. (2-1) to avoid unstable error propagation, a phenomenon peculiar to the beam (and plate) analyses because they involve second order difference equations. (See Ref 1).

The Table I class of mixed structural systems is selected for study because it simplifies development of coefficients in perfect numbers. It facilitates evaluation of the effects of matrix order, sparcity, number of calculations and relative stiffness, on manipulation errors. Since, as shown in Ref. 2, the completely full regular mixed structural system cannot be numerically singular, it admits evaluation of manipulation errors when equation sort is non-critical.

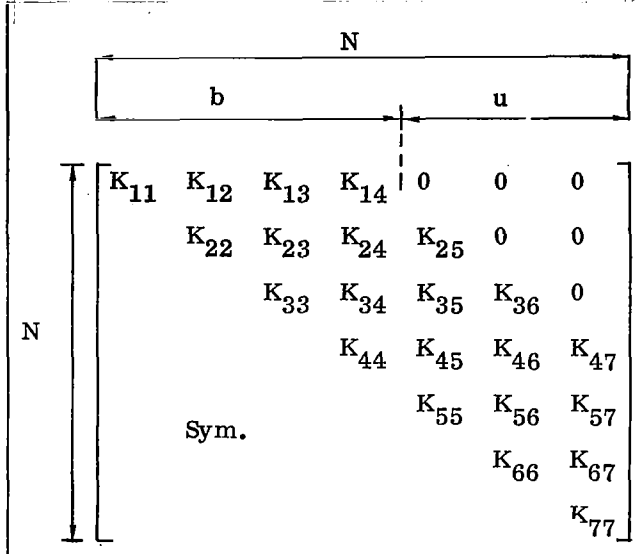
Solution Method

Policies for solving the load-deflection equations are as follows:

1. Coefficients of the stiffness matrix will be express in perfect numbers (integers).
2. The equations will be treated in the same sequence as they are expressed in Table I with boundary conditions imposed in the last equations.
3. Displacement boundary conditions will consist of clamping a single joint. Loadings will consist of reinforcing loads of equal value, at every joint. Two loadings will be used. In the first, every component of the load vector will be 1.0. In the second, every component will be $(2^{28}-1)/2^{27}$. This loading is the "bound" loading.
4. All structural analysis calculations will be performed in single precision using a computer with a 27 bit mantissa.
5. The mode of arithmetic will be truncation.

TABLE I

MIXED STRUCTURAL SYSTEMS



$$\begin{bmatrix} K_{11} & K_{12} & K_{13} & K_{14} & 0 & 0 & 0 \\ & K_{22} & K_{23} & K_{24} & K_{25} & 0 & 0 \\ & & K_{33} & K_{34} & K_{35} & K_{36} & 0 \\ & & & K_{44} & K_{45} & K_{46} & K_{47} \\ & & & & K_{55} & K_{56} & K_{57} \\ & & & & & K_{66} & K_{67} \\ & & & & & & K_{77} \end{bmatrix}$$

N = matrix order

u = uncoupled size

b = semibandwidth = N - u

i = row number

r = geometric progression ratio

$\begin{bmatrix} k_{11} & k_{12} \\ k_{12}^T & k_{22} \end{bmatrix}$ = element stiffness

$K_{11} = k_{11}(b-1)$

$K_{22} = k_{22} + (b-1)k_{11}r$

$K_{33} = (1+r)k_{22} + (b-1)k_{11}r^2$

$K_{44} = (1 + r + r^2)k_{22} + (b-1)k_{11}r^3$

$K_{55} = (r+r^2+r^3)k_{22} + (b-2)k_{11}r^4$

$K_{66} = (r^2+r^3+r^4)k_{22} + (b-3)k_{11}r^5$

$K_{77} = (r^3+r^4+r^5)k_{22}$

$K_{12} = K_{13} = K_{14} = k_{12}$

$K_{23} = K_{24} = K_{25} = rk_{12}$

$K_{34} = K_{35} = K_{36} = r^2k_{12}$

$K_{45} = K_{46} = K_{47} = r^3k_{12}$

$K_{56} = K_{57} = r^4k_{12}$

$K_{67} = r^5k_{12}$

Therefore $K_{ii} = f_1k_{11} + f_2k_{22}$ and $K_{ij} = k_{12}r^{i-1}$

where $f_1 = (b_i - 1)r^{i-1}$ if $b_i > 1$

$$f_2 = \begin{cases} 0 & \text{if } i = 1 \\ \dfrac{r^{i-2}(1 - r^{\eta})}{r^{\eta-1}(1 - r)} & \text{if } i > 1,\ r \neq 1 \\ \eta & \text{if } i > 1,\ r = 1 \end{cases}$$

with $\eta = \begin{cases} i - 1 & \text{if } i-1 \leq b-1 \\ b - 1 & \text{if } i-1 > b-1 \end{cases}$

The first policy avoids random inherited errors. The second is consistent with the Ref. 2 results. These show that joint sequencing based on minimizing wavefront is usually sufficient and sequencing from the most flexible to least flexible region is best. The third and fifth policies are selected to maximize manipulation errors. Reference 2 reports that manipulation errors reduce as additional kinematic constraints are imposed while Ref. 1 shows substitution errors are maximized by using reinforcing loading. Truncation is used since it incurs larger standard deviations of error than rounding. Twenty-seven bit single precision is selected to complement comparable precision test data cited in Ref. 2. Reference 1 shows errors vary inversely with the precision, so results of tests reported here can be projected to any other binary computer.

Manipulation Error Measurements

Error measures include evaluations for decomposition, numerical singularity, diagonal division, substitutions, and total closure. These evaluations are performed in a precision higher than for the solution precision. Double-precision was used here in evaluating error measures. The error measures are described in the paragraphs that follow.

Decomposition Error Measure.- The decomposition error is measured by:

$$e_D = \left(E_{ii} / K_{ii}\right)_{MAX.} \quad i = 1, 2, \ldots N \tag{2-8}$$

where e_D is the decomposition relative error,
N is the matrix order,
E_{ii} is the diagonal element of the matrix $K - L D^{-1} L^T$, and
K_{ii} is the corresponding diagonal of K.

Singularity Error Measure.- This error is measured by:

$$e_S = \frac{b^{1-p}}{r_S} \tag{2-9}$$

where e_S is the singularity relative error, and
r_S is the minimum (D_{ii} / K_{ii}) where D_{ii} is the i^{th} diagonal of D.
p is the precision and
b is the base (2.)

This measure is modified from that given in Ref. 2 to reflect maximum errors rather than expected.

Substitution Error Measure.- Forward, diagonal, and backward substitutions involve solving equations of the form,

$$A y = c \tag{2-10}$$

where A is any of the decomposition factors L, L^T, or D^{-1},
x is the vector of unknowns, and
c is the known right hand side.
Then, the substitution error measure is given by

$$e_B = \frac{\| Ay - c \|}{\| c \|} \tag{2-11}$$

where e_B is the substitution relative error, and
$\| \ \|$ is the norm "operator".

Note that Eq. (2-11) defines error measures involving residuals.

Measures of substitution error are taken for forward, e_{B1}, diagonal, e_{B2}, and backward e_{B3} substitutions. Both Euclidean norms and the norm which is the sum of the absolute value of vector components are used.

A second error measure for evaluating back substitution errors is the direct back substitution error measure. It is defined by

$$e^D_{B3} = \frac{\| \psi_{iD} - \psi_{iS} \|}{\| \psi_{iD} \|} \qquad (2\text{-}12)$$

where e^D_{B3} is the higher-precision error measure for back substitution relative error,
ψ_{iD} is the value of ψ_i developed using higher-precision arithmetic (p=54), and
ψ_{iS} is the value of ψ_i obtained using basic problem precision arithmetic (p=27).

Total Error Measure.- This error is measured by evaluating,

$$e_T = \frac{\| K\Delta - F \|}{\| F \|} \qquad (2\text{-}13)$$

where e_T is the total relative error.
This measure like the substitution measures of Equation (2-11) measures the imbalance in the equations when the calculated solution is introduced into the original equations.

Equation (2-13) provides a residual measure of total error. An alternate measure is developed by considering the elastic work. The work error is measured by

$$e_W = \frac{\Delta^T F - \Delta^T K \Delta}{\Delta^T F} \qquad (2\text{-}14)$$

where e_W is the work relative error.
The first term in the numerator is external work and the second, the internal work. Note that this error measure is signed. The negative sign indicates that internal work exceeds external.

The work error measures total error by evaluating the internal work using Eq. (2-3) and (2-6), i.e.,

$$\Delta^T K \Delta = \Delta^T L^T D^{-1} L \Delta = z^T D^{-1} z \qquad (2\text{-}15)$$

Equation (2-14) transforms Equation (2-13) to the form

$$e_{TW} = \frac{\Delta^T F - z^T D^{-1} z}{\Delta^T F} \qquad (2\text{-}16)$$

where e_{TW} measures the total work error.
The relation between the total error and component errors can be developed by considering the interaction of errors in each operation. Suppose the error in the solution is such that

$$\Delta = K^{-1}(I + e_T)F \tag{2-17}$$

The corresponding decomposition can be written in similar form as

$$L^T D^{-1} L = K(I + e_D) \tag{2-18}$$

Then the solution, using the decomposition algorithm, can be expressed by

$$\begin{aligned}\Delta &= L^{-1}(I + e_{B3})\, D^{-1}(I + e_{B2})\, L^{T-1}(I + e_{B1})(I + e_D)F \\ &\approx K^{-1}(I + e_D + e_{B1} + e_{B2} + e_{B3})F\end{aligned} \tag{2-19}$$

where terms involving products of the error matrices have been assumed negligible and dropped in the expansion. Then, comparing Eq. (3-3) and (3-5) and taking the matrix norms,

$$e_T \leq \|e_\Sigma\| \tag{2-20}$$

with $$e_\Sigma = \|e_D\| + \|e_{B1}\| + \|e_{B2}\| + \|e_{B3}\|$$

e_Σ is called the summed error measure.

Number of Calculations.- To relate error to the number of calculations, the formulas given in Table II are required. These define the calculations involved in each of the equation solving operations in terms of the parameters defined in Table I. The number of calculations in forward and back substitution are the same.

In applying these formulas to matrices of the class defined in Table I, X varies by one in alternate rows for beam elements. Therefore, X is taken as the average bandwidth (wavefront).

These error measures are selected based on the experience reported in Ref. 1 and 2. Only equation solution errors are checked because input output errors are negligible, and coefficient generation errors are usually small, controllable, and easily sensed. The decomposition error, Equation (2-8) was found adequate, though it requires $2N(W+1)$ calculations, where W is the wavefront. The singularity error measure was found to be the most important and efficient measure for sensing poor calculation sequencing. The substitution measures are introduced to facilitate numerical evaluation of the relative importance of error sources. The direct back substitution measure is used to determine the importance of error measure definition. The work measure is proposed as an efficient measure which could be included in production programs. Residual error measures are selected because they provide exacting measures of error which can be used to estimate errors in stress predictions. These errors were shown to be unbounded in displacement analyses.

The data developed in evaluating these error measures permits determining the magnitude of errors as a function of the number of calculations and equations. Reference (2), pages 58-65 and page 101, shows error bounds based on these parameters are usually one to two orders of magnitude high for displacement analyses and one to four orders high for force. Tests for both methods show that the bounds are realizable in computer analyses.

TABLE II

Formulae for Number of Calculations

Operation	Multiplications	Additions
Decomposition	$u\left(\frac{b^2}{2}+\frac{b}{2}-1\right)+\frac{b}{2}\left(\frac{b^2}{3}+b-\frac{4}{3}\right)\approx 2N^3d^2$	$\frac{ub}{2}(b-1)+\frac{b}{6}(b^2-1)\approx 2N^3d^2$
Division	$(u+b)=N$	0
Substitution	$2ub+b^2+b\approx 2N^2d$	$2ub+b^2-2u-b\approx 2N^2d$

Table III

Subcritical Test Problem Sizes*

Problem Number	No. of Equations	Density %	Calculations Decomposition	Substitutions	Total
1	50	13.0	5.63^{2}	6.06^{2}	1.17^{3}
2	100	6.75	1.13^{3}	1.21^{3}	2.34^{3}
3	200	3.43	2.25^{3}	2.40^{3}	4.65^{3}
4	200	5.37	5.81^{4}	3.99^{4}	9.80^{3}
5	200	20.2	8.68^{5}	1.61^{4}	1.03^{5}
6	200	28.7	1.80^{5}	2.31^{4}	2.03^{5}
7	200	36.9	3.01^{5}	2.98^{4}	3.31^{5}
8	200	44.5	4.46^{5}	3.60^{4}	4.82^{5}
9	200	51.6	6.11^{3}	4.18^{3}	6.52^{4}
10	300	3.61	8.74^{4}	5.99^{4}	1.47^{5}
11	300	11.3	8.88^{5}	2.00^{4}	1.09^{5}
12	300	17.4	2.15^{5}	3.12^{4}	2.46^{5}
13	300	23.3	3.90^{5}	4.20^{4}	4.32^{5}
14	300	29.0	6.11^{5}	5.24^{3}	6.63^{5}
15	400	1.73	4.50^{4}	4.80^{3}	9.30^{3}
16	400	2.72	1.17^{4}	8.00^{4}	1.97^{4}
17	400	7.56	9.38^{5}	2.37^{4}	1.18^{5}
18	400	11.4	2.13^{5}	3.59^{4}	2.49^{5}
19	400	15.0	3.78^{5}	4.79^{4}	4.26^{5}
20	500	2.18	1.46^{4}	1.00^{4}	2.46^{4}
21	500	6.85	1.50^{5}	3.36^{4}	1.83^{5}
22	500	9.90	3.15^{5}	4.90^{4}	3.64^{5}
23	600	1.16	6.75^{3}	7.20^{3}	1.40^{4}
24	600	1.82	1.75^{4}	1.20^{4}	2.95^{4}
25	600	3.13	5.32^{4}	2.15^{4}	7.47^{4}
26	600	5.73	1.80^{5}	4.04^{4}	2.21^{5}
27	700	0.99	7.88^{3}	8.40^{3}	1.63^{4}
28	700	1.56	2.04^{4}	1.40^{4}	3.44^{4}
29	700	2.69	6.21^{4}	2.51^{4}	8.72^{4}
30	800	0.87	9.00^{3}	9.60^{3}	1.86^{4}
31	800	1.37	2.34^{4}	1.60^{4}	3.94^{4}
32	1000	0.70	1.13^{4}	1.20^{4}	2.33^{4}

*Exponents indicate a power of ten. e.g., $0.1^{4} = 0.1 \times 10^{4}$.

Section 3

TESTS OF SUBCRITICAL SIZE PROBLEMS

This section describes experiments and their errors for problems involving less than 13.4×10^6 calculations. These data give values for all the error measures described in Section 2, as a function of matrix order, the number of calculations, and matrix sparsity.

Subcritical Test Problems

Table III cites the set of 32 subcritical test problems. The table lists equation order, density, and the number of calculations involved in solution. These problems vary from 200 to 1000 order. They include from 606 to 3.94×10^4 calculations.

The table below cites similar data for equations associated with the analysis of steerable antenna reflectors. Comparing data for these problems with those in Table IV suggests that the test problems have typical characteristics. The trend toward decreasing matrix density with increasing problem size signals the tendency for bandwidth to have a limiting value independent of problem size.

Table IV

Equations for Steerable Reflectors

Reflector Size	No. of Equations	Population Density
40 feet	105	0.270
60 feet	183	0.205
85 feet	282	0.180

The test problems were solved with a computer program which would work with the stiffness matrix in-core. The sequence of calculations in decomposing the row was as follows:

1. Form the reciprocal of the diagonal.
2. Multiply all elements in the row by the reciprocal.
3. For each element with higher row and column number than the diagonal, perform the subtraction

$$l_{jk} = l_{jk} - \frac{l_{ij} l_{ki}}{l_{ii}} \qquad (3\text{-}1)$$

where $j = i, i+1, \ldots N$; the row number of the element being adjusted
$k = i+1, i+2, \ldots N$, the column number of the element being adjusted.

These steps were repeated as i varied from one to N, the matrix order. These operations produce coefficients of the L matrix with coefficients of D on the diagonal.

The remaining solution steps are as follows:

4. Evaluate each y_j in turn by subtracting components of the inner product from the loading vector component.

5. Divide each y_j by the corresponding D_j to evaluate z_j
6. Evaluate each x_j in turn by subtracting components of the inner product vector from the z_j component.

Test Results

Table V lists values of the measured errors for the 32 test problems This table cites error for the bound loading. (See item 3 of solution method on page 4). It cites Euclidean norms of the error vectors.

There is little difference in errors for these two loading conditions. Over the 32 problems the difference in measured errors between loadings with components of 1.0 and the bound loading is generally less than 1.6%. Each loading gives maximum total error in about half of the runs.

The selection of the types of norm, on the other hand, has an important effect of the magnitude of indicated errors. The ratio of the Euclidean to the absolute norms varies from 1.13 to 1.31 over the problem set for the total error measure. This suggests that at least 18% deviation from an indicated trend can arise if different norm measures are taken. The same range of ratios of errors as a function of norm is also observed for substitution errors.

Interpreting Error Data

There is a redundancy of data in Table V. Two independent measures of total error are provided by e_T and e_{TW}. A third total error measure is reduced by applying Eq. (2-19) to the data. Furthermore, e_{B3} and e^D_{B3} provide two independent measures of back substitution error. To reconcile the differences in magnitude of these comparable measures, the subset of problems representing cantilevered beams, problems 1, 2, 3, 15, 23, 27, and 30, will be examined in detail.

Figure 1 shows the relations between error in tip deflection prediction and the number of equations for these problems. The continuous curve portrays the exact manipulation error for the bound loadings. The long-dashed curve shows how the energy error varies with order. This curve is plotted directly from data in Table V.

The exact error is developed by comparing the analytic solution of the difference equations with the computer predicted tip deflections. The analytic solution expresses tip deflections as

$$12\, u_{z_{N-1}} = \frac{Q a^4}{EI}\left(3J^4 + 12J^3 + 7J^2 - 2J\right) \qquad (3\text{-}2)$$

where $J = N/2$. The evaluations of Eq. (3-2) and the associated tip deflections and errors are summarized in Table VI. These data show the bound loading has larger errors than that involving perfect loading. This is due to the persistence of perfect numbers through forward substitutions with the unit loading and is reflected by the zero errors in Table V.

Both of the curves in Fig. 1 show a monotonic increase in relative error as the number of equations increases. The exact error is about ten times the indicated energy error throughout the range. Thus the energy error measure correlates with the actual errors in displacements.

Table V

Test Results for Subcritical Problems

Problem Number	Decomp $\left(\frac{E_{ii}}{K_{ii}}\right)_{MIN.}$	Fwd. $\frac{\|Ly-F\|}{\|F\|}$	Div. $\frac{\|D^{-1}z-y\|}{\|y\|}$	Bkwd. $\frac{\|L^T\Delta-z\|}{\|z\|}$	Dbl. Prec. $\frac{\|\nu_{iD}-\nu_{is}\|}{\|\nu_{iD}\|}$	Total $\frac{\|K\Delta-F\|}{\|F\|}$	Work $\frac{\Delta^TF-z^TD^{-1}z}{\Delta^TF}$
	e_D	e_{B1}	e_{B2}	e_{B3}	e_{B3}^D	e_T	e_{TW}
1	0	1.56^{-7}	5.02^{-10}	1.77^{-6}	9.03^{-9}	2.55^{-3}	3.25^{-8}
2	0	3.11^{-7}	1.95^{-10}	8.07^{-6}	1.38^{-7}	4.37^{-2}	5.46^{-8}
3	0	6.36^{-7}	1.19^{-10}	3.12^{-5}	5.96^{-7}	4.78^{-1}	9.04^{-8}
4	2.24^{-8}	9.50^{-7}	6.56^{-9}	1.45^{-6}	5.94^{-7}	2.08^{-4}	1.22^{-9}
5	1.49^{-7}	6.34^{-7}	6.29^{-9}	1.03^{-6}	6.12^{-7}	2.32^{-5}	2.30^{-8}
6	2.15^{-7}	6.64^{-7}	7.33^{-9}	9.46^{-7}	6.25^{-7}	1.74^{-5}	3.18^{-8}
7	2.87^{-7}	7.21^{-7}	6.36^{-9}	9.03^{-7}	6.44^{-7}	1.62^{-5}	7.88^{-7}
8	3.51^{-7}	7.75^{-7}	5.72^{-9}	8.78^{-7}	6.55^{-7}	1.64^{-5}	1.25^{-7}
9	4.09^{-7}	8.32^{-7}	8.14^{-9}	8.15^{-7}	6.62^{-7}	1.65^{-5}	1.66^{-8}
10	2.24^{-8}	1.19^{-6}	7.17^{-9}	2.18^{-6}	9.22^{-7}	4.62^{-4}	-3.26^{-8}
11	1.13^{-7}	9.13^{-7}	6.29^{-9}	1.63^{-6}	8.79^{-7}	6.24^{-5}	-8.58^{-8}
12	1.81^{-7}	9.18^{-7}	6.03^{-9}	1.52^{-6}	8.83^{-7}	3.94^{-4}	-1.13^{-8}
13	2.59^{-7}	9.90^{-7}	5.72^{-9}	1.50^{-6}	9.34^{-7}	3.29^{-5}	4.14^{-8}
14	3.52^{-7}	1.04^{-6}	6.25^{-9}	1.48^{-6}	9.76^{-7}	3.04^{-5}	6.76^{-8}
15	0	1.27^{-6}	6.52^{-11}	1.36^{-4}	1.17^{-6}	1.13^{-1}	1.99^{-7}
16	2.01^{-8}	1.85^{-6}	7.01^{-9}	2.86^{-6}	1.23^{-6}	8.16^{-4}	2.42^{-8}
17	8.02^{-8}	1.18^{-6}	6.29^{-9}	2.05^{-6}	1.07^{-6}	1.16^{-4}	-5.52^{-8}
18	1.53^{-7}	1.18^{-6}	5.79^{-9}	2.02^{-6}	1.10^{-6}	7.50^{-5}	-3.61^{-8}
19	2.41^{-7}	1.23^{-6}	6.44^{-9}	1.97^{-6}	1.15^{-6}	5.83^{-5}	-1.87^{-9}
20	2.24^{-8}	2.28^{-6}	6.14^{-9}	3.92^{-6}	1.59^{-6}	1.43^{-3}	-4.79^{-8}
21	1.08^{-7}	1.51^{-6}	6.01^{-9}	2.59^{-6}	1.33^{-6}	1.59^{-4}	-6.66^{-8}
22	1.83^{-7}	1.47^{-6}	5.62^{-9}	2.64^{-6}	1.38^{-6}	1.12^{-4}	-7.39^{-8}
23	0	1.97^{-6}	3.51^{-11}	2.85^{-4}	1.72^{-6}	5.41^{1}	3.30^{-7}
24	2.03^{-8}	2.78^{-6}	7.01^{-9}	4.40^{-6}	1.83^{-6}	1.90^{-3}	-4.22^{-7}
25	5.19^{-8}	2.09^{-6}	6.71^{-9}	3.43^{-6}	1.60^{-6}	5.40^{-4}	-2.11^{-8}
26	1.08^{-7}	1.83^{-6}	6.94^{-9}	3.20^{-6}	1.60^{-6}	2.34^{-4}	-7.41^{-8}
27	0	2.33^{-6}	2.39^{-11}	4.25^{-4}	1.86^{-6}	1.08^{2}	5.09^{-7}
28	2.24^{-8}	3.15^{-6}	6.81^{-9}	5.56^{-6}	2.17^{-6}	2.79^{-3}	-1.03^{-8}
29	5.20^{-8}	2.44^{-6}	6.33^{-9}	4.16^{-6}	1.83^{-6}	7.72^{-4}	2.22^{-8}
30	0	2.56^{-6}	1.71^{-11}	5.38^{-4}	2.22^{-6}	1.77^{2}	5.05^{-7}
31	2.24^{-8}	3.73^{-6}	6.43^{-9}	5.84^{-6}	2.47^{-6}	3.34^{-3}	-6.50^{-9}
32	0	2.85^{-6}	9.83^{-12}	7.91^{-4}	2.96^{-6}	4.09^{2}	4.57^{-7}

*Exponent implies a base of 10. e.g., $.1^{-4} = .1 \times 10^{-4}$

° $e_s = 4.48^{-8}$ for all problems except 1, 2, 3, 15, 23, 27, 30, and 32. For these, $e_s = 9.96^{-8}$. $e_s = b^{1-p}/r_3$

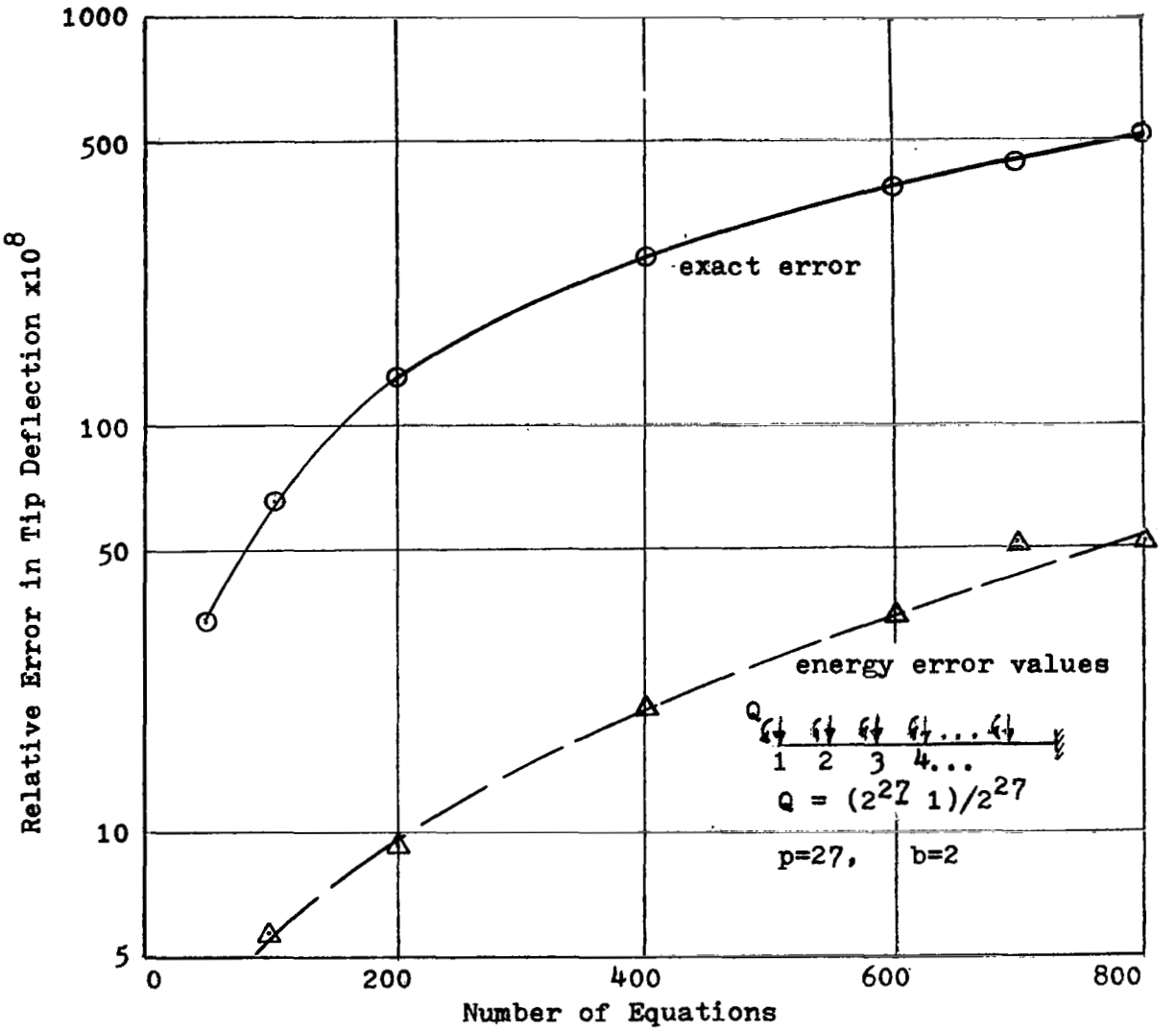


Figure 1. Exact Displacement and Energy Error Growth

Table VI

Cantilevered Beam Analysis Errors*o

Problem Number	No. of Equations	Tip Defl. Eq. (3-2)	Q ≐ 1.0 e	Q = Bound e	e_T
1	50	0.11364167^{6}	5.88^{-8}	3.18^{-7}	2.55^{-3}
2	100	0.16889500^{7}	5.92^{-7}	6.46^{-7}	4.37^{-2}
3	200	0.26005817^{8}	1.41^{-7}	1.32^{-6}	4.78^{-1}
15	400	0.40802330^{9}	7.27^{-7}	2.62^{-6}	1.13^{1}
23	600	0.20520525^{10}	—	3.88^{-6}	5.41^{1}
27	700	0.37945088^{10}	—	4.37^{-6}	1.08^{2}
30	800	0.64640932^{10}	1.58^{-6}	5.23^{-6}	1.77^{2}
32	1000	0.15750146^{11}	—	6.58^{-6}	4.09^{3}

*Powers imply a ten base. e. g., $3.^{-7} = 3. \times 10^{-7}$.

oEquations in optimum sort, b = 2, p = 27.

The last column of Table VI repeats selected data from Table V. Comparing these data with errors listed in the next to last column shows the total residual error measurements also exhibit monotonic error growth. However, error magnitudes are four to nine orders of magnitude greater than actual deflection errors and vary over the problem range.

An explanation for this hypersensitivity is exposed by examining the process of error evaluation. Multiplication of the stiffness matrix by the displacements is intrinsically a differencing operation. As deflections increase, this incurs critical arithmetic. For example, in problem 2 deflections of the order of 10^7 must be differenced to develop load components of the order of 10^0. Therefore seven digits of accuracy are lost in the process. Since only 8.2 digits are carried when p = 27, the accuracy of is a maximum of 1.2 digits. Since the last digit of the deflections involves truncation error, the residual error is non-zero. Its exponent is dependent on the exponent for the deflections and the calculation precision. The exponent of the error to the base 10 cannot be less than about 8.2 less than the deflection exponent, which has little to do with analysis error. Thus the total residual error does not reflect the accuracy of deflection predictions and cannot be expected to correlate with it.

These data show that large residual errors may disguise a relatively accurate analysis (6 digit accuracy for the 8 digit precision calculation in problem 2). On the other hand, small residual errors can be expected to correlate with solution accuracy. This is illustrated by comparing the errors indicated by the two back substitution error measures e_{B3} and $e_{B\delta}^{D}$.

Figure 2 facilitates this comparison. It displays plots of back substitution residual and solution errors for two sets of problems from Table V. The continuous curves pertain to the cantilevered beam problems. The dashed curves are for problems 4, 10, 16, 24, 28, and 31. Since the solution error measure is based on double precision analysis it must be more accurate in predicting error than the residual. For the cantilevered beam problems poor correlation is obtained between this measure and the error indicated by the residual. This poor correlation is attributed to the critical arithmetic involved in the evaluation of the residual error measure. When the sparsity of the stiffness matrix is decreased (as for the dashed curve cases) good correlation is obtained between the error measures. Even these curves exhibit a tendency to lose their parallelism as the number of equations increases. Since an increase in the number of equations results in increasing deflections, this deterioration of the residual measure echoes the increasing criticality of arithmetic in the error evaluation.

Because of the infidelity of the residual error measures, the sum of the errors does not compare with the total error. This disagreement with Eq. (2-19) is due to the large error in evaluating e_T. Despite this disagreement, the residual measures are expected to indicate errors for other than the cantilevered beam problems in these tests. For the cantilevers, they provide exaggerated estimates of errors.

The cantilevered beam problems can also be used to illustrate the relation between manipulation and discretization error. Consider that the beams are loaded with a uniform pressure $q=\frac{Q}{a}$, couples of value Q at each joint, and a tip load of magnitude $qa/2$. Then the tip deflection is given by

$$12u_{3N-1} = \frac{Qa^4}{EI}(3J^4 + 12J^3) \qquad (3\text{-}3)$$

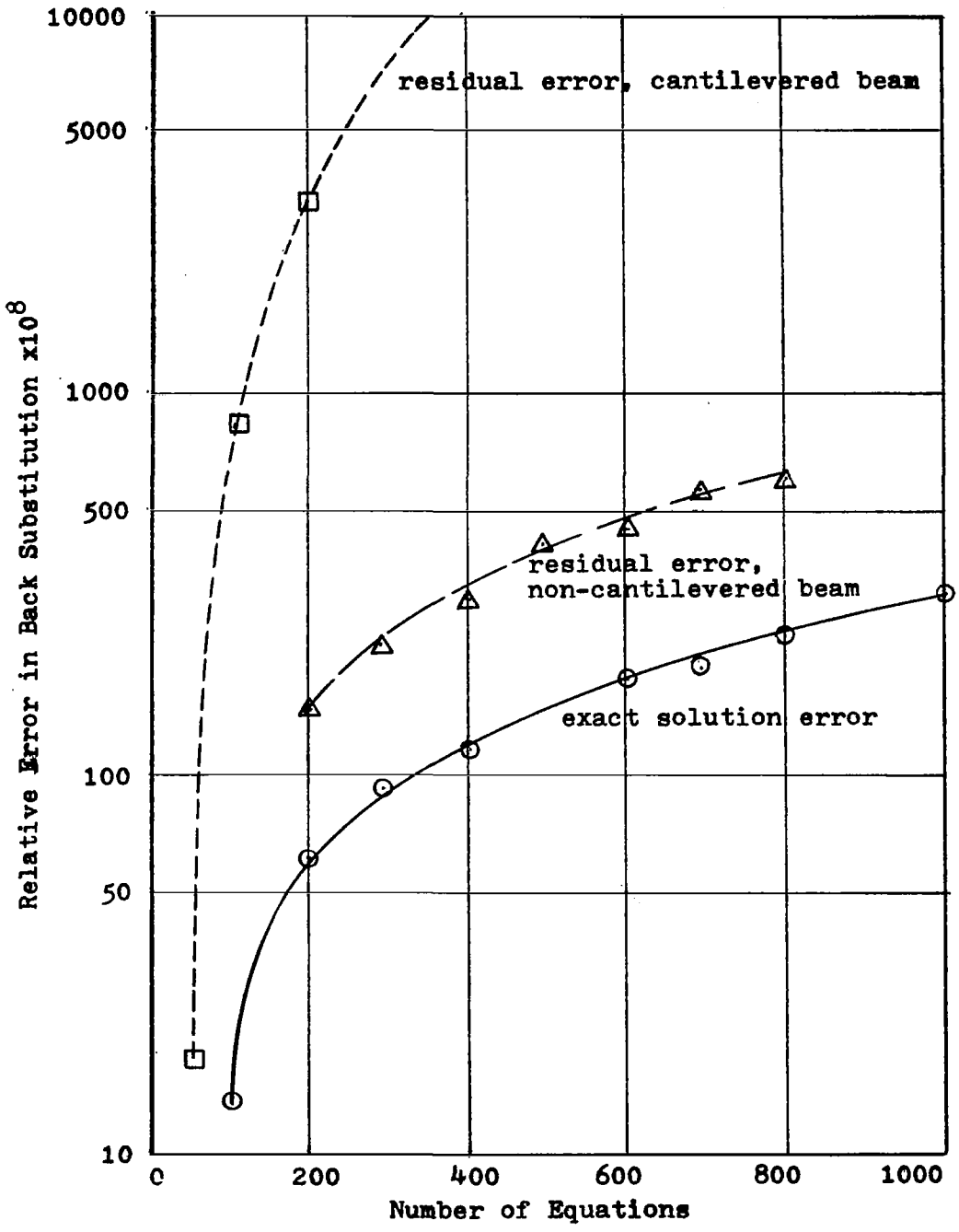


Figure 2. Residual and Solution Error Growth

The difference between Eqs. (3-2) and (3-3) determines the discretization error. This difference is never zero, regardless of how large N becomes. The relative discretization error, however, diminishes to zero monotonically as J increases to infinity.

Figure 3 provides a plot showing the total, manipulation, and discretization error for these beams. The total error is depicted by the continuous curve. This is obtained by comparing computer generated deflection predictions with those of Eq. (3-3). The manipulation errors are those cited in Table VI. The discretization error is found from Eqs. (3-3) and (3-2).

The total error reflects a monotonic decrease of discretization error and monotonic increase of manipulation error as the number of equations increases. The interaction of the error components produces a total error curve of quadratic form. For these problems, the manipulation error reduces tip deflection predictions and discretization provides overestimates. Thus, the total error is expected to be zero for two problem sizes. The curve of total error shows both of these occur at more than 1000 equations. Discretization error dominates the total error trend until the number of equations exceeds 800.

Error Magnitudes

To interpret error magnitudes, their values will be compared with the maximum errors associated with that simple series subtraction used in Ref. 1, page 21. The subtraction operation consists of performing a number of subtractions, N_c , such that the result of each subtraction yields an answer which is opposite in sign to the next component to be subtracted.

Reference 1, Section 2, shows that the error in this operation, when the worst number representation is used, is greater than for any other "simple arithmetic" operation. Moreover, it has a larger error bound than vector multiplication. Study of this operation identifies an error bound. Since this bound is both dependent on the sequence of arithmetic and the number representation, it is expected to define an upper bound for error for computer calculations, independent of the analysis involved.

Reference 1, page 21, concludes that the error must be less than

$$e_m = b^{-p} N_c \qquad (3\text{-}4)$$

where e_m is the maximum relative error, and

Nc is the number calculations.

Equation (3-4) is an approximation in the region where error growth is maximum. When $N_c < b^{p-2}$, the critical number of calculations, the error may exceed that given by Eq. (3-4). When $N_c > b^p$, the error bound will be less than that of Eq. (3-4). Equation (3-4) is an approximation in the region where error growth is maximum. Errors from each error source will be compared with those of Eq. (3-4).

Decomposition Diagonal Errors (e_D)

Figure 4 is a log-log plot of the decomposition error for the 24 subcritical test problems with non-zero error. Each continuous curve applies to a particular number of equations. Dashed curves define contours of equal matrix density.

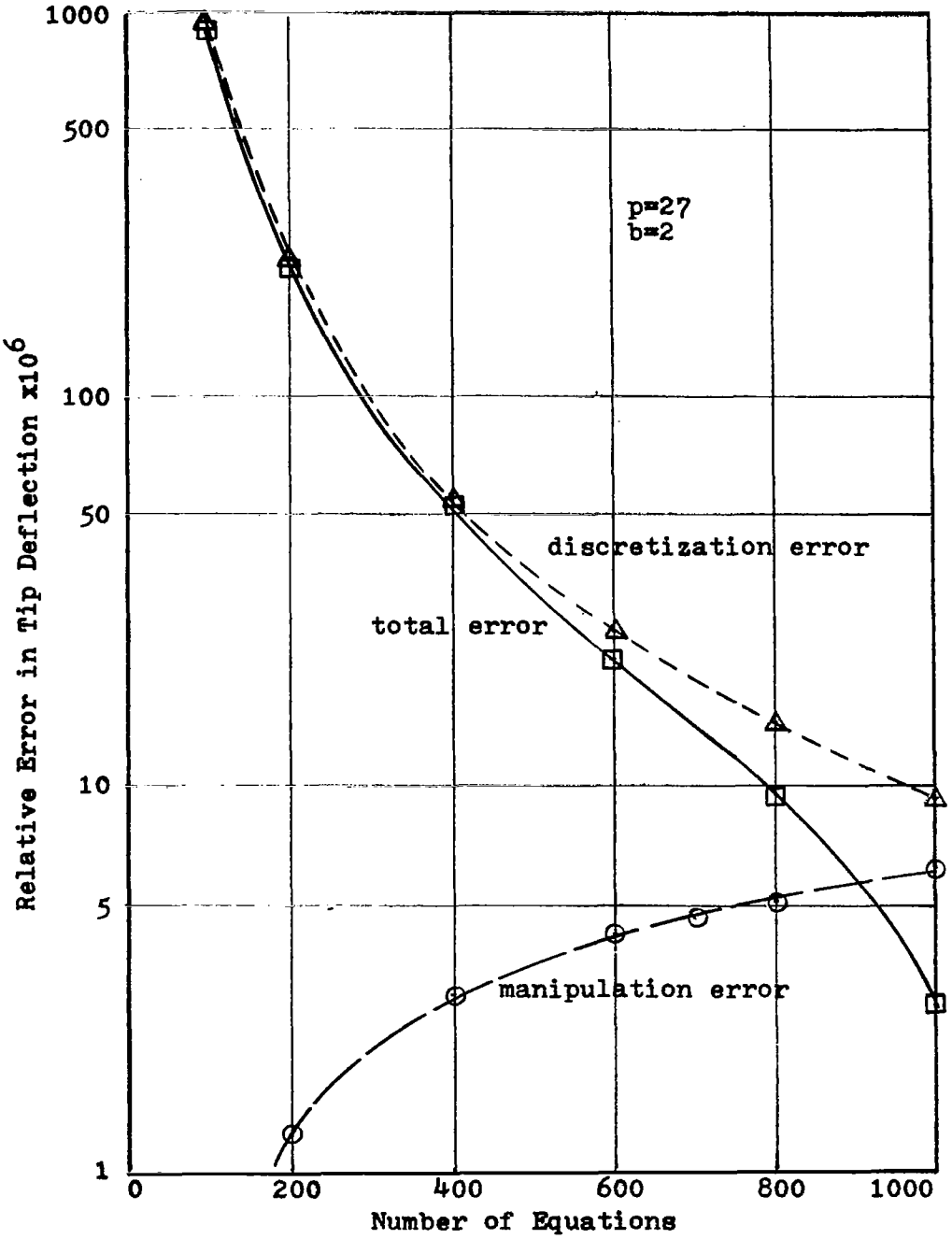


Figure 3. Manipulation and Discretization Error Growth

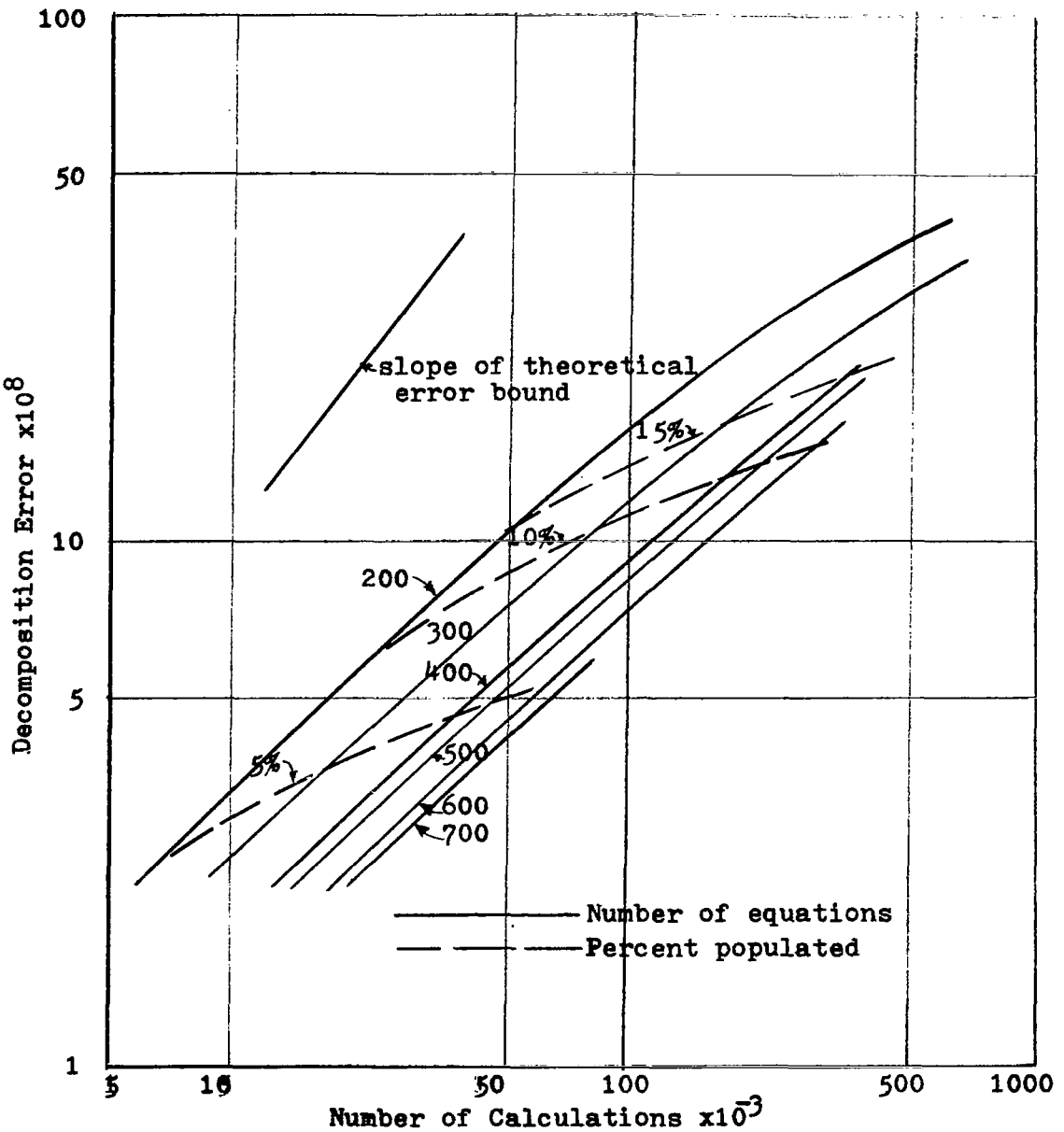


Figure 4. Decomposition Error Trends

These data show decomposition diagonal errors:

1. Increase with increases in the number of calculations for a given matrix order,
2. Increases for a given matrix density as matrix order increases, and
3. Decrease for a given number of calculations as matrix order increases.

The short line segment shows the slope that the curves would take if the error were directly proportional to the number of calculations. The slope of both the continuous and dashed curves are always less than the segment slope. The error magnitudes indicated in Fig. 4 are 10^{-5} of those given by Eq. (3-4) for these subcritical problems.

Forward Substitution Errors (e_{B1}).- Figure 5 is a log-log plot of the forward substitution errors for the 32 subcritical problems. Each continuous curve applies to a particular number of equations. Dashed curves define contours of equal matrix density. The circled points designate cantilevered beam problems.

These data show forward substitution errors:

1. Decrease then increase with increases in the number of calculations for a given matrix order,
2. Increase for a given matrix density as matrix order increases, and
3. Increase for a given number of calculations as matrix order increases.

Indicated error magnitudes are less than 10^{-5} of those given by Eq. (3-4). Note that the cantilevered beams display smaller errors than other problems.

Figure 6 provides plots of forward substitution errors against the number of equations. Continuous curves connect problems with the same matrix joint bandwidth. One set of problems consists of the cantilevered beams (joint bandwidth = 2). The second set includes errors of problems 4, 16, 24, 28, and 31. (Joint band = 3). These data show that though the substitution errors increase linearly with matrix order for the cantilevered beams the growth is more rapid when the bandwidth is larger.

Diagonal Division Errors (e_{B2} and E_s).- Since the maximum relative error in division is b^{-p+1}, the maximum value for the diagonal division error for these tests is 1.492×10^{-8}. Assuming it is equally likely that the error is zero or one in the last bit, the expected error is 7.46×10^{-9}.

The 24 experiments reported in Table V (omitting the cantilevered beam cases) exhibit an average diagonal division error of 6.18^{-9}. Considering both sets of loading conditions, the range of the error varies from 5.61^{-9} to 8.14^{-9}. No concerted trend for errors to increase with matrix order occurs, or would be expected. Similarly, error magnitudes are independent of matrix population density.

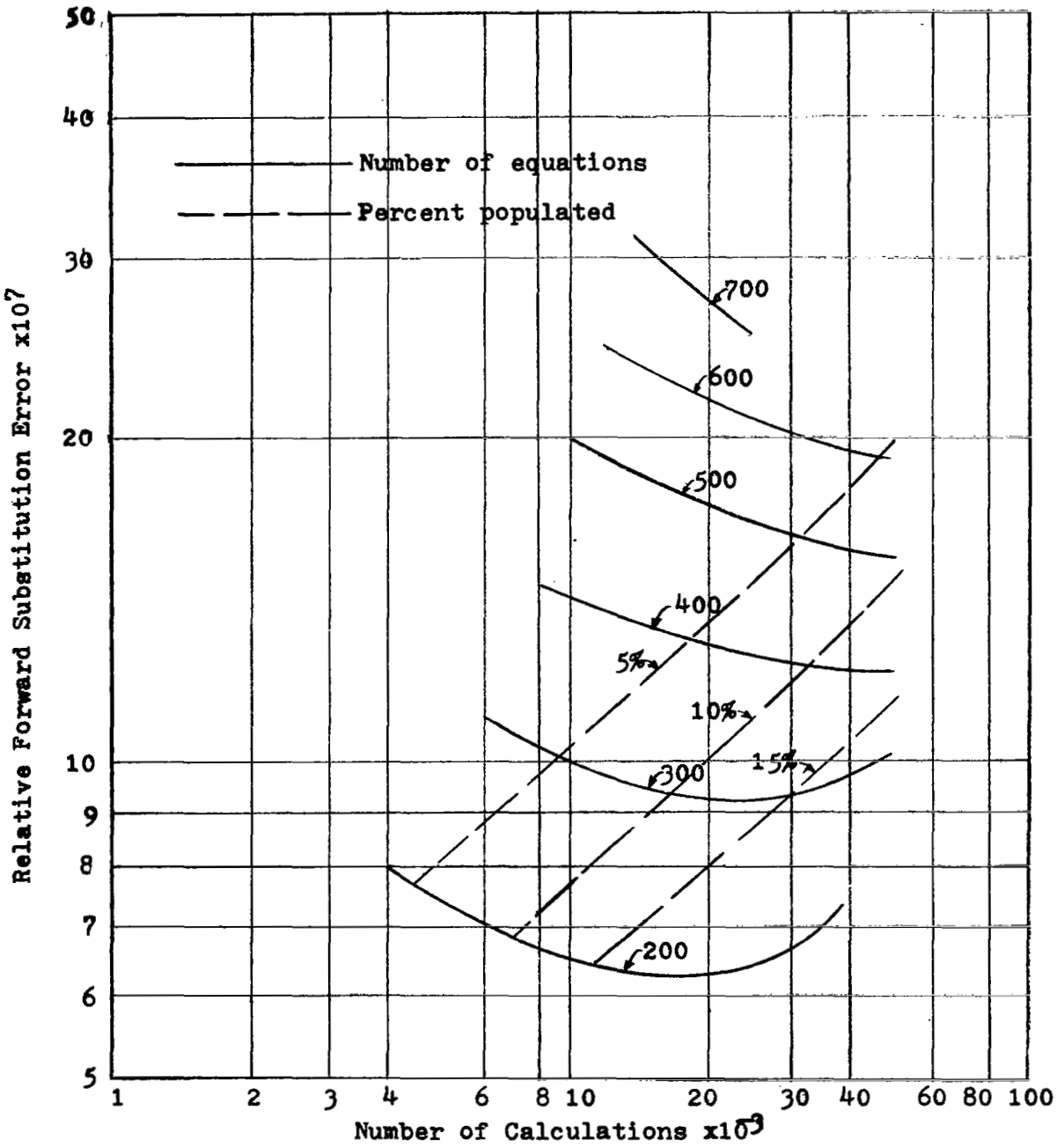


Figure 5. Forward Substitution Error Trends

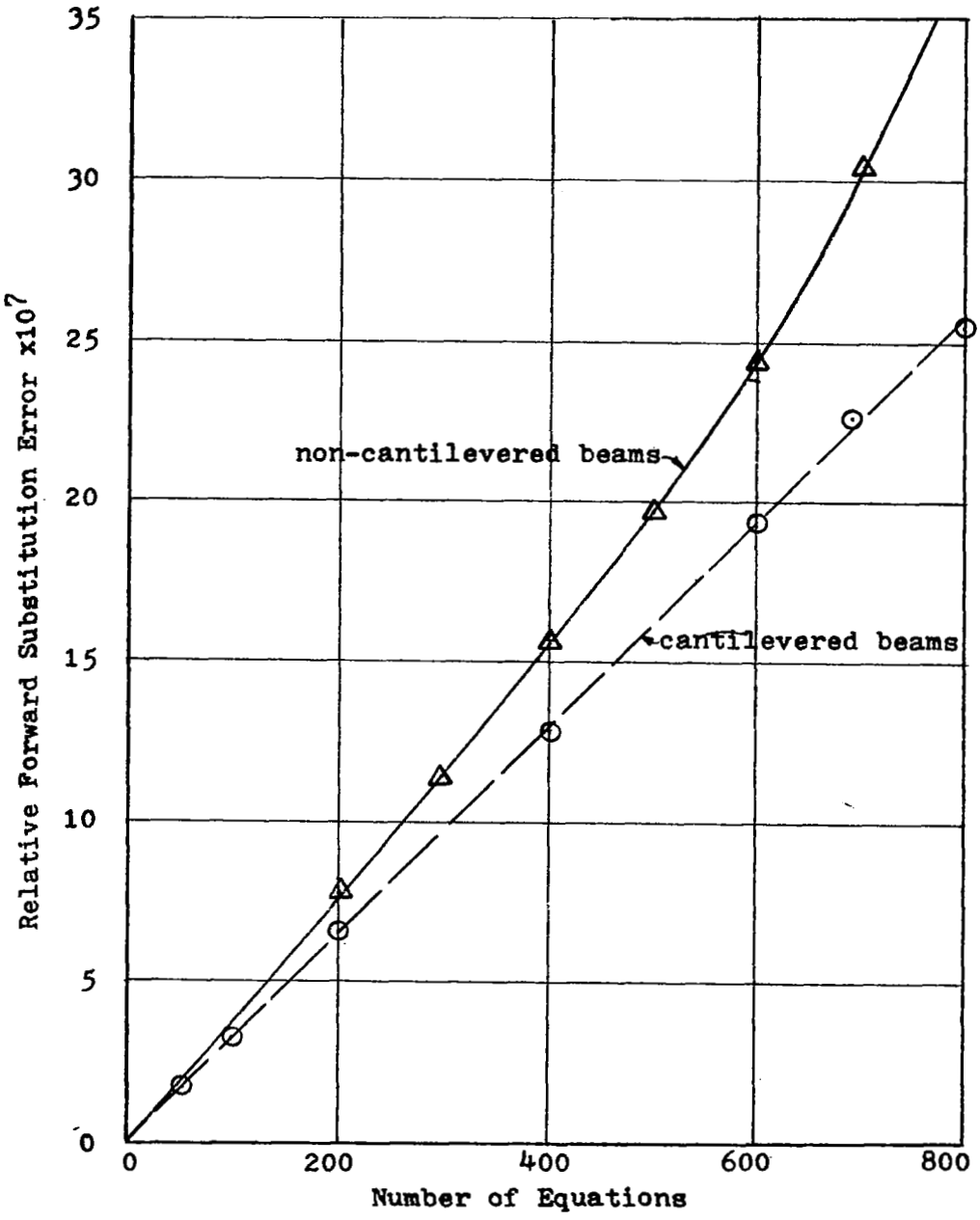


Figure 6. Forward Substitution Errors for Constant Band Matrices

The set of problems representing cantilevered beams show division errors that are much smaller than the average. Moreover, the errors decrease monotonically with problem size. This occurs because half the diagonals are perfect numbers and result in no residuals though the denominator norm is increased nevertheless. Since this is unrepresentative, these errors are disregarded in forming the average and noting trends above.

For all problems, singularity ratios had a minimum value of 0.125. This confirms the adequacy of equation sequencing in the analysis. It indicates the loss of a maximum of the last three binary places in the mantissa due to numerical singularity. Thus, the maximum relative error is the determinant of the stiffness matrix, and hence in the answers, due to numerical singularity is 0.996×10^{-7}. The expected error is half this amount.

Back Substitution Errors (e_{B1R}).- Figure 7 is a log-log plot of the back substitution errors for the 32 subcritical problems. Each continuous curve applies to a particular number of equations. Dashed lines define contours of equal matrix density. Residual errors associated with cantilevered beam problems are two orders of magnitude greater than those for the other problems, for the same number of calculations, so these results for cantilevers do not appear on the graph.

These data show back substitution errors:

1. Decrease with increase in the number of calculations for a given matrix order,
2. Increase for a given matrix density as matrix order increases, and
3. Increase for a given number of calculations as matrix order increases.

Error magnitudes indicated in Fig. 7 are 10^{-3} of those given by Eq. (3-4).

Relative Importance of Error Sources.- The data in Table V indicates that the sources of error in order of decreasing error magnitude are back substitution, forward substitution, decomposition, and division errors. This ordering is deceptive, however, because the implications of a given magnitude error depend on error source. Moreover, this evaluation omits consideration of inherited errors. An examination of the cantilevered beam problems can provide some of the desired perspective.

Figure 8 shows the tip deflection error, due to each error source, as a function of the number of equations for the set of cantilevered beam problems.

The displacement error due to inherited error is a plot of

$$e_i = -0.229 N^2 \gamma \qquad (3\text{-}5)$$

where e_i is the relative error in tip deflection,
N is the number of equations,
γ is the relative error in the diagonal stiffness matrix coefficients. $\gamma = 2^{-27}$

This equation is developed from data in Table VI of Ref. 1. It is assumed that every diagonal of the stiffness matrix has an error of one part in the last binary position due to addition of element stiffness matrices.

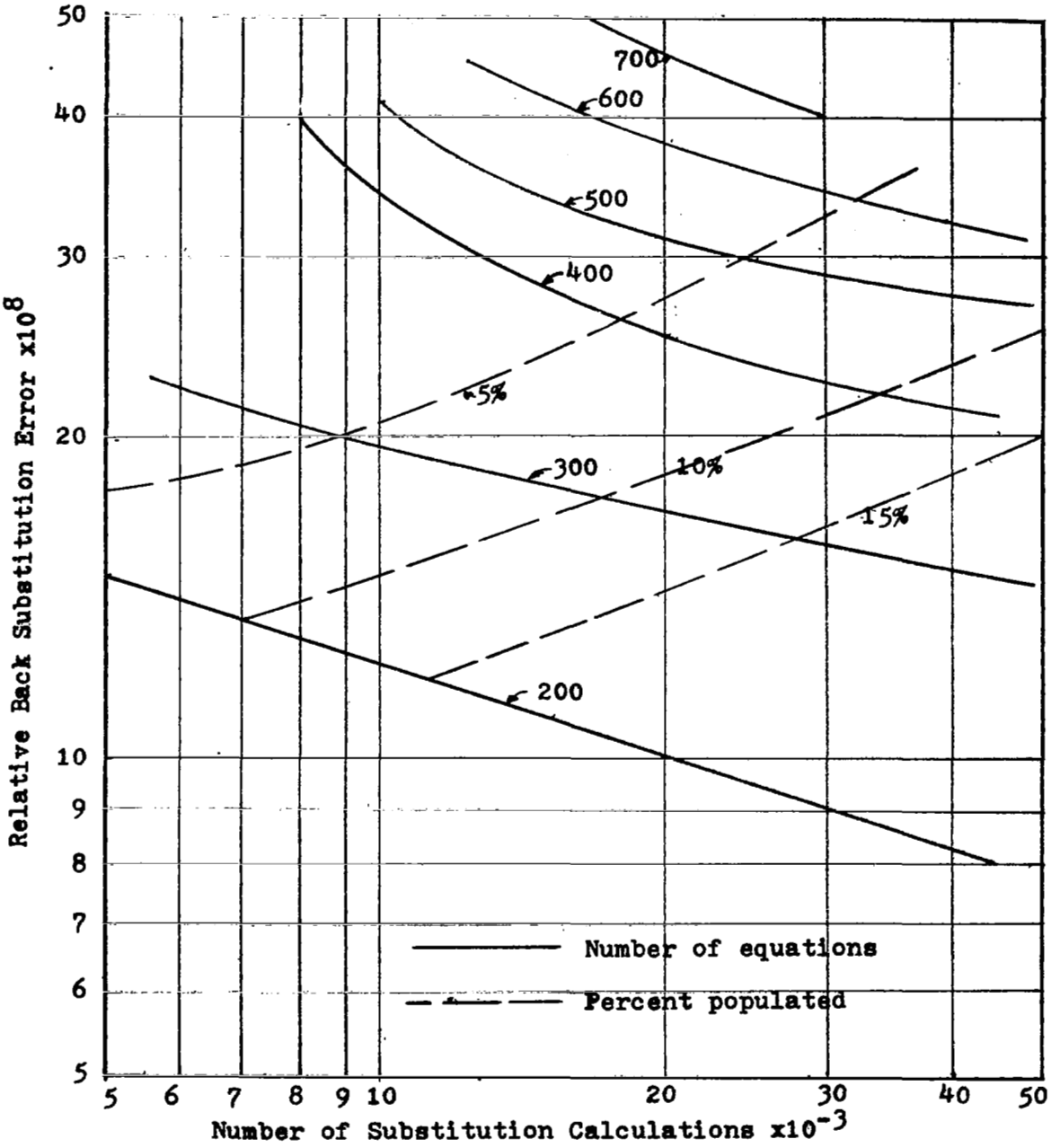


Figure 7. Back Substitution Error Trends

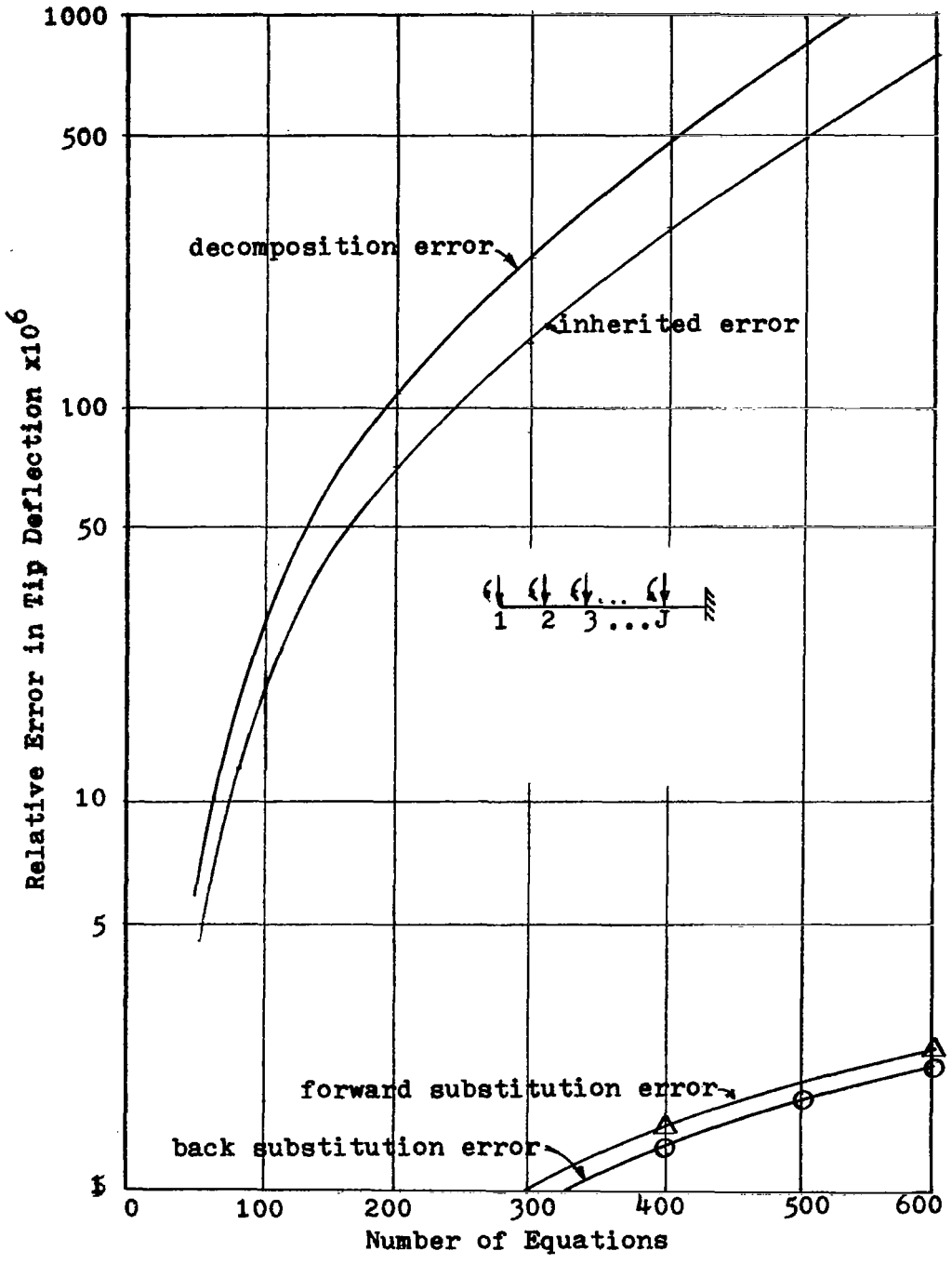


Figure 8. Equation Solution Error Magnitudes

The displacement error due to decomposition error is also defined by the function given by Eq. (3-5). In this case, however, γ is taken as the maximum value of decomposition error measured in the 32 tests: $\gamma = .224 \times 10^{-7}$. (This value is more realistic than the zero error of cantilevered beams).

Since the forward substitution error implies an error in the loading and the equations are linear, the consequent error in tip deflections are taken to be equal to the measured forward substitution errors. These result in the lowest curve on the figure.

The effect of diagonal division error is to simulate an error in the determinate of the stiffness matrix. Consequently, these errors also are direct measurements of implied solution errors. These errors, however, are so small that they do not appear in the range plotted.

Back substitution errors directly effect predicted displacements. Thus these solution errors, as measured, are plotted for the comparison.

Figure 8 shows that the relative importance of errors based on solution implications is, in decreasing order of importance, decomposition, inherited, back substitution, forward substitution, and diagonal division. Decomposition and inherited errors are much more important than substitution errors. If inherited errors are dominant, Eq. (3-4) (which is based on a simple beam problem) may be used to furnish a guide on the relation of error and precision.

Section 4

TESTS OF CRITICAL SIZE PROBLEMS

Tests to evaluate the significance of performing more than b^P calculations require the use of a different computer code. This section describes results of additional tests to evaluate the effect of changing details of the solution process to facilitate critical problem solution and error results for a set of problems involving from 55×10^4 to 40×10^6 calculations.

Modified Analysis Approach

Efficient data handling requires modifying the sequence of arithmetic for solving large problems using the algorithm defined by Eqs. (2-1) through (2-6). Details of the data handling for the critical size problems is described in Ref. 3. The following changes to the sequence of performing the calculations described in Section 2 are involved:

1. During decomposition, sum all subtractive terms before making a single subtraction to modify the ℓ_{jk} element for all elements in rows up to row j.

2. During decomposition do not divide elements in the row by the reciprocal of the diagonal.

3. No changes.

4. and 5. Solve directly for z_j by subtracting components of the inner product from the loading component and dividing the result by the diagonal.

6. Solve for each x_i by subtracting components of the inner product vector from the z_j component and multiply the result by D_j^{-1}.

These minor changes result in no increase in the number of non-trivial calculations. They involve only minor changes in the sequence of arithmetic. The change represented by 1, above, would be expected to reduce manipulation error. The changes represented by steps 2 through 6 may increase, decrease, or not affect error.

Manipulation Error Measurements

Only two measures are used in the critical size problems. The first measure is the numerical singularity test. The second measure is a solution test based on comparing the calculated deflections with the exact solution of the computer difference equations. This measure takes the form

$$e_\Delta = \frac{\| \Delta_{iE} - \Delta_{iC} \|}{\| \Delta_{iE} \|} \qquad (4\text{-}1)$$

where e_Δ is the relative deflection error,
Δ_i is a deflection component, and
the subscript E means "exact" and the subscript C, "calculated."

The Euclidean norm is used as a measure of the numerator and denominator of Eq. (4-1) for most test problems. For cantilevered beam problems, the maximum error component (tip deflection) is used herein.

For most problems, a new analysis procedure was adopted so the exact solution would be known. This process involved the following steps:

1. Assume the displacements in the form of perfect numbers. For the analyses three sets of displacements are assumed. The first, Δ involves unit displacements for every Δi; the second, Δ involves unit displacements in odd numbered equations only, and the third, Δ, has unit displacements in even numbered (moment) equations only.

2. Multiply the stiffness matrix times the assumed displacements. All the stiffness coefficients and assumed displacements are perfect numbers. In addition, each component of the multiplication solution vector will have a value less than 134,217,727 (the maximum number represented in the mantissa when p = 27, b =2). Therefore, the product will be taken with zero error.

3. Solve the load-deflection equations using the loadings from step 2.

4. Calculate the error using the vectors of step 1 as the exact solution and those of step 3 as the calculated.

Critical Test Problems and Results

The two sets of test problems and test measurements are summarized in Tables VII and VIII.

Table VII cites problem parameters and error measurements for the set of problems to evaluate the effect of the changes in the solution process. These problems represent uniform cantilevered beams with various numbers of joints and loadings. Data in rows 1 through 5 of this table define the magnitude of the numerical analysis problem. Data in rows 6 through 9 relate to analysis of these systems using the analysis methods described in Section 2. Rows 10 through 13 cite results using the manipulation error measurements and analysis methods just described.

Table VIII lists data for mixed system tests to evaluate errors when the number of calculations approaches and exceeds b^p, the critical number. Again rows 1 to 5 cite problem size data. Rows 6 through 8 list measured relative deflection errors for the three loadings of interest.

Error Magnitudes

Analysis of data in rows 6 through 9 of Table VII shows that minor changes in the arithmetic sequence can have a significant effect on error magnitudes. Row 8 gives the tip deflection relative error using the modified analysis approach. Row 9 lists the same type errors using the analyses approach described on page 4. Comparison of these two rows of data shows the modified approach has from 1.33 to 1.40 times the error of the standard. For these problems, decomposition is

Table VII

Solution Process Test Problems

Row No.	Problem	A	B	C	D	E	F
1	No. of Equations	600	700	800	1000	1600	2000
2	Density, %	1.16	0.99	0.87	0.70	0.44	0.35
3	N_c, decomposition	6.75^{3}	7.88^{3}	9.00^{3}	1.13^{4}	1.80^{4}	2.25^{4}
4	N_c, substitutions	7.20^{3}	8.41^{3}	9.61^{3}	1.20^{4}	1.92^{4}	2.40^{4}
5	N_c, total	1.39^{4}	1.62^{4}	1.86^{4}	2.33^{4}	3.72^{4}	4.60^{4}
6	Exact Defl. Eq. (3-2)	2.05205245^{9}	3.79450877^{9}	6.46409314^{9}	1.57501458^{10}	1.02912373^{11}	2.5100583^{11}
7	Calc. Defl.	2.05204144	3.79448570	6.46404833^{9}	1.57500073^{10}	1.02911070^{10}	2.50996478^{11}
8	e_x	5.42^{-6}	6.08^{-6}	6.93^{-6}	8.78^{-6}	1.27^{-5}	3.74^{-5}
9	e_x(Table V)	3.88^{-6}	4.37^{-6}	5.23^{-6}	6.58^{-6}		
10	Density, %	1.82	1.56	1.37			
11	e	2.88^{-3}	4.05^{-3}	5.11^{-2}			
12	e	4.02^{-3}	5.44^{-3}	7.08^{-2}			
13	e	7.27^{-5}	9.83^{-5}	1.25^{-4}			

*Exponents imply a base of ten, e.g. $8.^{-6} = 8.0 \times 10^{-6}$

Table VIII

Critical Size Test Problems

Row No.	Problem	G	H	I	J	K	L	M	N
1	No. of Equations	200	300	400	400	500	600	800	1200
2	Density, %	5.16	29.0	15.0	72.5	9.90	3.13	42.0	29.3
3	N_c, decomposition	6.11^{5}	6.11^{5}	3.78^{5}	1.01^{7}	3.15^{5}	5.32^{4}	2.47^{7}	3.94^{7}
4	N_c, substitutions	4.18^{4}	5.24^{4}	4.79^{4}	2.34^{5}	4.89^{4}	2.15^{4}	5.40^{5}	8.45^{5}
5	N_c, total	6.53^{5}	6.63^{5}	4.26^{5}	1.03^{7}	3.64^{5}	5.54^{4}	2.52^{7}	4.02^{7}
6	e_{Δ}	8.06^{-6}	1.61^{-5}	3.55^{-5}	4.23^{-5}	6.74^{-5}	4.44^{-4}	1.08^{-4}	1.31^{-4}
7	e_{Δ}	2.02^{-5}	3.92^{-5}	8.56^{-5}	1.10^{-4}	1.54^{-4}	8.72^{-4}	2.91^{-4}	2.51^{-4}
8	e_{Δ}	8.74^{-6}	1.66^{-5}	3.52^{-5}	5.03^{-5}	5.90^{-5}	7.73^{-4}	1.40^{-4}	1.20^{-4}

is exact. Therefore the difference in solution details involve only those in the substitutions. It is concluded that an algorithm must consistently show more than a fifty percent reduction in error to be considered a significant improvement.

Comparing the data in rows 8 and 11 through 13 of Table VII with that in row 8 leads to three additional conclusions:

1. Error magnitudes are sensitive to the choice of loading. This sensitivity is reflected by a maximum factor of 41.3 between errors in the last three rows.

2. The relative error for two loadings cannot be added to predict the error for the sum of the loadings. If this were possible, e_{Δ} would be equal to $e_{\Delta} + e_{\Delta}$. This result reflects the nonlinearity of error with loading condition.

3. The cantilevered beam reinforcing loadings evokes relatively large errors. Comparing errors of row 6 and row 11 in Table VII shows the reinforcing loading (row 6) incurs a minimum of a factor of times the error of the worst alternate load. This result confirms a conclusion of Ref. 1.

An analysis of data in Table VIII confirms the first two conclusions above. Row 7 displays errors that are a maximum of 2.7 times those in either row 6 or 8. The sum of the row 7 and 8 errors does not equal the measured row 6 error.

Figure 9 displays the measured errors as a function of the number of calculations. The bound given by Eq. (3-4) is also shown. This plot shows measured errors are several orders of magnitude below the error bound. There is no indication that errors increase dramatically when the critical number of calculations is exceeded.

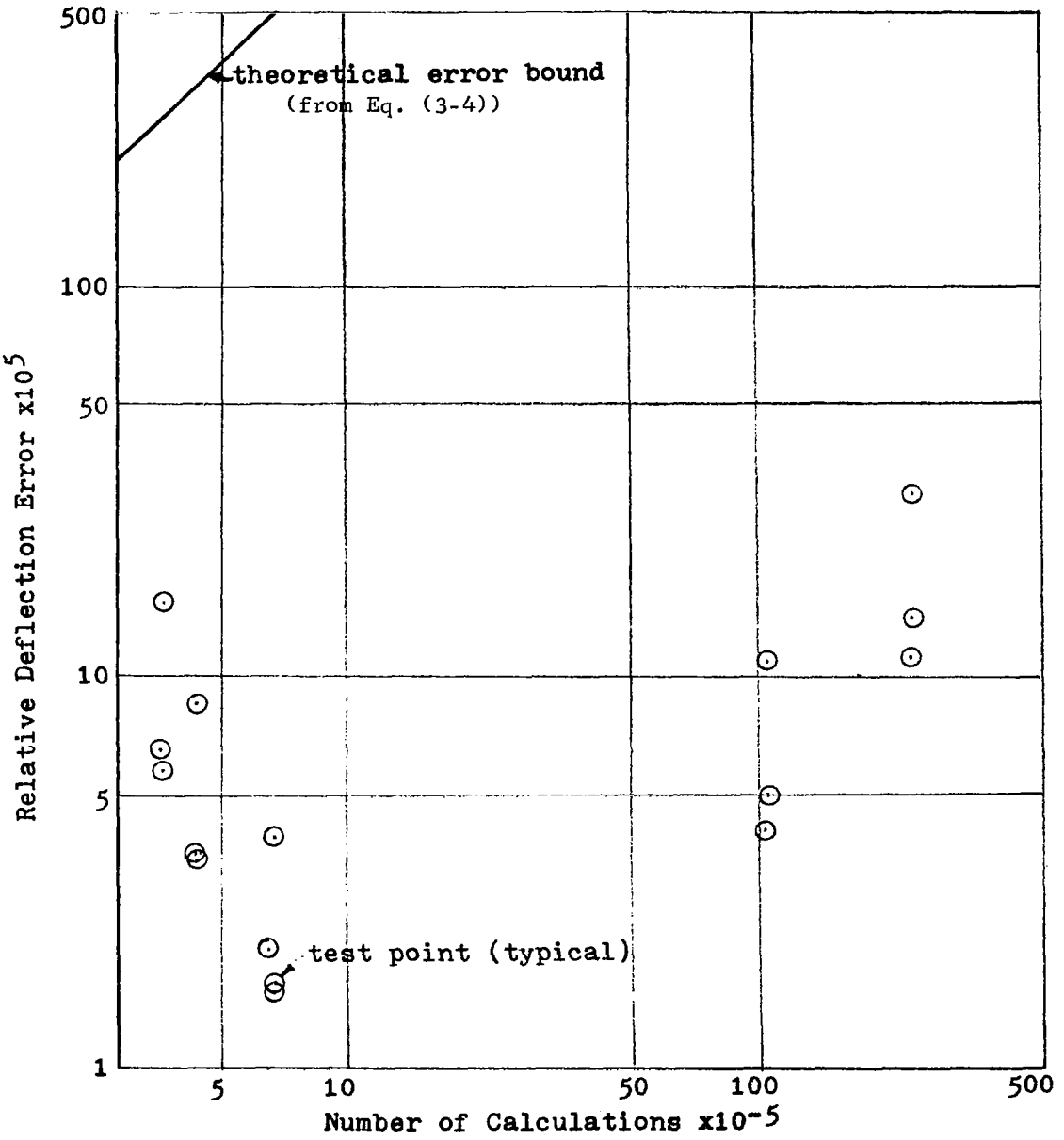


Figure 9. Deflection Errors for Critical Test Problems

Section 5

PRODUCTION CODE ANALYSIS CHECKS

The error measures used in this study were defined to discriminate among error sources and to study problems of error measurement for manipulation errors. Data from the test problems have identified the adequacy of the measures. In this section particular measures are recommended and others rejected for use in a production computer code for structural analyses. All of these checks should involve special coding to add them to an existing computer code. Use of matrix manipulation instructions to evaluate the measures will result in extra passes of data through core and thus incur unnecessarily large analyses time penalties on the computer.

Recommended Measures

Table IX lists the recommended error measures in the order in which they would arise in the usual analysis. Each of these checks requires few calculations compared with those of the solution process. The set of checks is intended to check calculations in critical errors and provide a measure of the accuracy of displacement and stress predictions.

The "number of calculations check" is proposed to eventually avoid equation solution if the probability of success is low. The idea is to include this calculation in all analyses in order to obtain statistical data relating problem size to indicated deflection error and computer precision. For this purpose, the number of calculations can probably be satisfactorily estimated by

$$N_c = 2\left(\sum_{i=1,2}^{N} f_i W_i^2 + 2\sum_{i=1,2,\ldots}^{N} f_i W_i\right) \qquad (5\text{-}1)$$

where f_i is the number of degrees of freedom at a joint i
W_i is the number of joints, with higher joint number than i, which are elastically coupled to joint i during the decomposition of row i.

W_1 is easily evaluated from the topology of the structure. The first term in Eq. (5-1) estimates the total number of decomposition calculations and the second, substitution calculations. The first term in Eq. (5-1) estimates non-trival calculations in decomposition. The second estimates calculations in substitution.

The numerical singularity check is recommended to insure that critical arithmetic does not destroy accuracy. Reference 3, page 42-43, shows that if this test indicates that large errors may arise, resequencing of the equations can usually eliminate the difficulty.

The positive definite check provides an overall check on the reality of the stiffness model. If any of the diagonals, D_{ii}, are less than zero it is implied that at least one structural deformation pattern can occur in which the principle of conservation of energy is violated. This may not identify an unacceptable mathematical model (for some geometries the Hrennikoff lattice exhibits this deficiency[4]) but it signals the possibility of unrealizable structural response predictions. If the diagonals are relative zeros, this test will identify the existence of kinematic instability and may be extended to differentiate between active and passive instabilities[5].

Table IX

Recommended Error Measures

Measure	Function	How Evaluated	When Evaluated	No. of Calcs.
Number of Calculations	Indicate problem size and desirability of cutting run	Eq. (5-1)	During preprocessing	4N
Numerical Singularity	Sense critical arithmetic during decomposition	Eq. (2-8)	During decomposition	2N
Positive Definiteness	Insure that deformations require work	Test $D_{i,i} \geq 0$	During back substitution	N
Energy Error	Evaluate norm of error predicted	Eq. (2-16)	During back substitution	5N+1
Stress Precision	Define error in stress	Eq. (5-2)	When stresses are evaluated	6N

The energy error measure provides a low cost measure of total analysis (deflection) error. This study has shown that it correlates with analysis error while avoiding critical arithmetic. At the same time, it reflects the loading of interest and the associated structural response. When incorporated in the back substitution process, it requires no special data handling or transfers.

The stress precision check, when used with the energy error measure, permits defining the relative accuracy of stress predictions. This check is defined as

$$e_{\sigma_i} = 1.0 - \frac{\|k_i \Delta_i\|}{\|k_i\| \cdot \|\Delta_i\|} \qquad i = 1, 2, \ldots E \qquad (5\text{-}2)$$

where e_{σ_i} is the relative stress precision error for element i,
k_i is the stiffness matrix for element i,
Δ_i is the subset of displacements for the joints of element i, and
E is the total number of elements.

If e_{σ_i} is near zero no significant error has been incurred in differencing deflections to evaluate stresses. If e_{σ_i} is near 1, as many digits have been lost in stress evaluation as there are nines following the decimal point. The total relative error in stress predicitions can be determined by adding e_{σ_i} to the relative energy error.

Consider again the cantilevered beam. Then Table X summarizes pertinent error data for four sets of equations. Each row of the table lists error data for one case. To illustrate the thinking in determining the accuracy of stress predictions, consider the second row of data. Calculations were performed with 8.3 digits precision (p = 27), so e_{TW} implies less than one digit loss of accuracy in calculating deflections. Since $e^{\circ}_{\sigma_{TIP}}$ is one to six digits, about six digits of precision are lost in stress predictions. Combining the error losses, the total loss is about six digits of accuracy. Therefore, of the 8.3 digits of precision, about two digits of accuracy remain.

The actual accuracy is listed in the second column of Table X. This is based on the number of digits of accuracy in estimate of the applied load found by taking the product of the stiffness matrix and the calculated deflections. Comparing the data in the second and last columns shows that accuracy predicted by Eq. (5-2) corresponds with the accuracy of element generalized force predictions.

Rejected Measures

Table XI cites several error measures whose use is rejected. In this table, costly measures are those which require as many calculations to evaluate as analysis of an additional loading. Very costly measures require as many calculations as the total equation solution process.

Tests reported here provide the basis for rejecting residual and solution measures. Reference 1 reports the inadequacy of Maxwell reciprocity tests. Reference 1 and Wilkinson[6] report the pessimism and unreliability of condition number measures.

Table X

Stress Measured and Predicted Accuracy*

N	Measured Digits in Tip Load Residual	Measured Error Parameters e°_{TW}	$\|\Delta\|^{\circ}$	$\|k_{TP}\Delta\|^{\circ}$	$e^{\circ}_{\sigma TIP}$	Predicted Digits of Accuracy
6	6	4.65^{-10}	2.00^{1}	2.00	$1\text{-}2.10^{-3}$	6
50	2	3.25^{-8}	1.56^{5}	2.00	$1\text{-}2.71^{-6}$	$2^{\$}$
100	1	5.46^{-8}	2.36^{6}	2.06	$1\text{-}1.84^{-7}$	1
200	0	9.04^{-8}	3.62^{7}	3.30	1-1.92-8	0

* Exponent implies a power of 10. e.g. $2.1^{-3} = 2.10 \times 10^{-3}$

° b = 2, p = 27, $\|k_{TP}\| = \sqrt{226.}$

\$ $e_{\sigma TIP}$ implies stresses depend on accuracy of last 8.3-5 ≈ 3 digits of deflections.

e_{TW} implies eighth digit of deflection is in error. Therefore, predicted accuracy is 2 digits.

Table XI

Rejected Error Measures

Measure	Basis for Rejection
Condition Numbers	• Very costly to evaluate • Lead to over conservative error bounds • Insensitive to loading of interest
Maxwell Reciprocity	• Costly to evaluate • Unreliable since it may not be an independent check
Reaction Check	• Unreliable since sample is too small • May incur critical arithmetic
Solution Error Check	• Very costly to evaluate when higher precision required
Total Residual Check	• Cost to evaluate • Inherently incurs critical arithmetic

Each of the measures recommended is reliable and requires few calculations. They are the better of the measures examined. Taken together they can identify critical numerical problem areas in the solution, define deflection and stress manipulation error magnitudes, and lead to statistical data relating the computer precision to accuracy.

Section 6

CONCLUSIONS

These tests have yielded the following conclusions on errors in numerical analysis of mixed beam systems:

1. The maximum error bound based on the number of calculations is very conservative. None of the tests gave decomposition errors closer than $\frac{1}{10}5$ of the bound, forward substitution errors closer than $\frac{1}{10}5$, backward substitution closer than $\frac{1}{10}3$. No evidence was educed showing a large increase in error occuring when b^p calculations are exceeded even though the bound increases at that level.

2. Indicated errors are sensitive to error measuring details, loading and details of arithmetic. Use of residual rather than solution error measures can make differences of two or more orders of magnitude in indicated errors. Change in loading can result in changes in error of at least one order of magnitude. Minor changes in arithmetic can change indicated errors by forty percent, even though the formula of the algorithm is unchanged. The errors measures are affected by about 25 percent by changing from the Euclidean to the absolute value norm.

3. Inherited and decomposition errors are the more important error sources. Forward and back substitution errors are relatively small and diagonal division errors negligible. This conclusion justifies use of the same precision arithmetic during decomposition as used in developing stiffness coefficients. A major observation is that higher precision need not be used in evaluating displacements unless stresses are required. Numerical singularity errors were also negligible in all the test problems indicating satisfactory equation sequencing.

4. Four error checks are recommended for numerical analysis of structures. The number of solution calculations should be estimated during preprocessing to yield statistical error data. Numerical singularity checks should be included in decomposition. Positive definite checks should be performed during diagonal division. The energy total error measure should be evaluated during back substitution. A stress precision check should be made as stresses are calculated. Rejected measures include Condition Number, reactions, total residual, Maxwell reciprocity, and direct solution error checks.

5. Discretization errors are much greater than manipulation errors even for beams modelled by up to 800 equations on a 27 bit mantissa computer. This is true when equations are in good sort (See Ref. 2, page 42.)

This study of manipulation errors in structural equation solutions has involved a semiempirical approach. Eight error measures will be evaluated for a range of problems to establish the relation between error magnitude and growth as a function of error source. These data have led to identification of four validated checks of displacement and stress prediction accuracy. The calculation penalties for these checks are negligible compared with the total solution calculations. These necessary and efficient checks are recommended for all codes involving computer numerical analyses of structures.

REFERENCES

1. Melosh, R. J. and Palacol, E. L.,"Manipulation Errors in Finite Element Analysis of Structures," NASA CR-1385, National Aeronautics and Space Administration, Wash., Aug. 1969, 141 p.

2. Melosh, R. J. and Palacol, E. L., "Manipulation Errors in Computer Solution of Structural Equations," NASA unpublished report to be released as a CR, Nov. 1969.

3. Melosh, R.J. and Bamford, R. M., "Efficient Solution of Load-Deflection Equations," Jour. Struct. Div. ASCE,Vol. ST4 No. 4, April 1969, p. 661-676.

4. Hrennikoff, A., "Solution of Problems in Elasticity by the Framework Method," J.A.M. Dec. 1941, pp A-169 - A-175.

5. Melosh, R. J., "Structural Analysis Frailty Evaluation and Redesign (SAFER), Technical Document," AFFDL-TR-70-15. Air Force Flight Dynamics Laboratory, Dayton, June 1970.

6. Wilkinson, J. H., Rounding Errors in Algebriac Processes, Prentice-Hall, Englewood Cliffs, N. J., 1963.