Numerical Comparisons of Some Explicit Runge–Kutta Pairs of Orders 4 Through 8

P. W. SHARP
University of Toronto

We compare the numerical performance of 6 explicit Runge–Kutta pairs of orders 4 through 8 on problems with continuous solutions. As part of this work we demonstrate new ways of presenting numerical comparisons. We first compare the efficiency of the pairs and the extent to which tolerance proportionality holds. Then we compare the accuracy of the local error estimate and stepsize prediction. Following this we compare the pairs on mildly stiff problems. Finally, we illustrate how the performance of the pairs can be affected by the selection of the stepsize.

Categories and Subject Descriptors: G.1.7 [**Numerical Analysis**]: Ordinary Differential Equations—*initial value problems*; G.4 [**Mathematics of Computing**]: Mathematical Software—*certification and testing*

General Terms: Algorithms, Performance, Reliability

Additional Key Words and Phrases: Efficiency explicit pairs, high order, low order, Runge–Kutta

1. INTRODUCTION

The nonstiff first-order initial value problem

$$\mathbf{y}' = \mathbf{f}(x,\mathbf{y}), \qquad x \in [x_0, x_f]$$
$$\mathbf{y}(x_0) = \mathbf{y}_0 \tag{1.1}$$

where $\mathbf{f}$: $\mathbf{R} \times \mathbf{R}^n \to \mathbf{R}^n$ is often solved numerically using an explicit Runge–Kutta pair. A pair generates approximations $\mathbf{y}_{i+1}$ and $\hat{\mathbf{y}}_{i+1}$ to $\mathbf{y}(x_{i+1})$, $i = 0, \ldots,$ according to

$$\mathbf{y}_{i+1} = \mathbf{y}_i + h \sum_{j=1}^{s} b_j \mathbf{f}_j \tag{1.2a}$$

$$\hat{\mathbf{y}}_{i+1} = \mathbf{y}_i + h \sum_{j=1}^{s} \hat{b}_j \mathbf{f}_j \tag{1.2b}$$

This work was supported by the Information Technology Research Centre of Ontario and the Natural Science and Engineering Research Council of Canada.
Author's address: Department of Mathematics, Queen's University, Kingston, K7L 3N6, Canada.

where (1.2a) is order p, (1.2b) is order $p-1$, s is the number of stages, $h = \mathbf{x}_{i+1} - \mathbf{x}_i$, and

$$\mathbf{f}_j = \mathbf{f}\left(x_i + c_j h, \mathbf{y}_i + h \sum_{k=1}^{j-1} a_{jk} \mathbf{f}_k \right), \qquad j = 1, \ldots, s.$$

Many classes of pairs have been derived. Fehlberg [6] derived (5, 6), (6, 7), (7, 8), and (8, 9) pairs of 8, 10, 13, and 17 stages, respectively. A German summary of the derivation for the (5, 6) and (7, 8) classes is given in Fehlberg [7]. Verner [14] derived pairs of the same order and the same number of stages as Fehlberg, except for his (8, 9) pairs which use only 16 stages. Fehlberg [8] gave lower order pairs including the well-known RKF45 pair. Dormand and Prince [1] derived classes of 6 and 7 stage (4, 5) pairs. The pairs in the latter class reuse the last stage as the first stage of the next step (FSAL). Later, Prince and Dormand [11] derived 8-stage (5, 6) and 13-stage (7, 8) pairs. A 5-stage (3, 4) pair due to Norsett is given in Enright et al. [4]. Other classes of pairs have been derived. However, we restrict our attention to pairs from the cited classes because these pairs have been used more widely, such as in numerical testing, or when investigating other properties of Runge–Kutta methods.

Classes of pairs generally have two or more coefficients as free parameters. Suitable values for these parameters are found by selecting sets of coefficients that satisfy one or more criteria. These criteria are intended to ensure that the pairs have desirable properties such as an accurate local error estimate. The selection of values for the free parameters is discussed in several papers (see Verner [13] or Prince and Dormand [11], for example).

The existence of more than one pair naturally leads to the questions of which pair to use, and whether this choice is affected by the type of problem being solved or the accuracy required. These questions are answered using two approaches. Numerical testing has shown that there is often a qualitative agreement between how well a pair satisfies the above criteria and its performance (see Prince and Dormand [11], for example). Hence, pairs can be compared using these criteria. However, this approach has at least two disadvantages. The effect of the higher order terms is generally not considered in these criteria. These terms may be significant at lax tolerances. Also, it is difficult to state criteria that enable pairs of different orders to be compared satisfactorily.

The second approach is to use numerical testing. A battery of test problems is solved for a range of tolerances, and the performance of the pairs is compared. This approach also has disadvantages. The performance of the pairs depends on the test problems used, how the pairs are implemented, and to a lesser extent, how the performance is evaluated. Despite these disadvantages, useful information can be obtained, as several studies have shown.

Enright and Hull [3] compared Fehlberg's (4, 5), (5, 6), (6, 7), (7, 8), and (8, 9) pairs with rational extrapolation methods, variable-order variable-step linear multistep methods, and Runge–Kutta methods that estimate the local

error using step halving. They found a lax tolerances that low order pairs are more efficient than high order pairs, while at severe tolerances the high order pairs are more efficient. A third conclusion was that a Runge-Kutta pair of the appropriate order is the most efficient of the methods tested if the derivative is inexpensive to evaluate. Verner [13] compared his (5, 6), (6, 7), (7, 8), and (8, 9) pairs with those of Fehlberg using the same test set. Verner found his pairs generally use more derivative evaluations, but they have a smaller maximum global error. They also take fewer steps with the true local error greater than the local error tolerance. Prince and Dormand [11], and Dormand and Prince [2] compared their (4, 5), (5, 6), and (7, 8) pairs with the pairs of Fehlberg and Verner. They found on nonlinear problems that their (7, 8) pair is often more efficient than the other pairs, even at lax tolerances. This appears to contradict the first conclusion of Enright and Hull. But both test sets compare individual pairs of a class, and conclusions reached about the pairs are not necessarily applicable to the whole class. Prince and Dormand also verified for mildly stiff problems that choosing formulas with large stability regions leads to a reduction in the number of derivative evaluations.

Although in the above numerical studies the quality of the local error estimate for the pairs is considered, the emphasis is on the efficiency of the pairs. In this paper we perform a more extensive numerical study of some explicit Runge-Kutta pairs. We compare the efficiency of the pairs, to what extent tolerance proportionality holds, the accuracy of the local error estimate and stepsize prediction, the performance on mildly stiff problems, and the effect of varying the strategy for selecting the stepsize. For the tests we used a modified version of DETEST (Enright and Pryce [5]). The modifications are described in Section 2.

Increasing the amount of testing naturally increases the amount of data that must be analyzed and presented. To reduce this to an acceptable amount, we limit our investigation to problems that have smooth solutions and to tolerances no more severe than 10^{-10}. We do not use CPU time to compare the pairs, and not all of the above pairs are tested. The first two restrictions mean we do not have to consider the effects of discontinuities, singularities, and roundoff error (the tests are done in 16-digit arithmetic). The third restriction means the results will not be influenced by our programming style.

We select pairs that use local extrapolation, because this is used in most integrators, and that are known from previous studies to be efficient or reliable among pairs of the same order. Also, we select more than one pair at some orders so that the effect of order can be eliminated. We test the (3, 4) pair of Nørsett, the (4, 5) pairs of Fehlberg [8] and Dormand and Prince [1], RK5(4)7FM], the (5, 6) pairs of Prince and Dormand [11] and Verner [14], and the [7, 8] pair of Prince and Dormand [11]. These pairs are denoted by N34, F45, DP45, V56, PD56, and PD78, respectively. Fehlberg gives several (4, 5) pairs in [8]. To avoid possible confusion we give the tableau of the pair we use in the appendix.

Despite the above limitations on the scope of our investigation, a considerable amount of data is generated. The second purpose of this paper is to show how this data can be presented in a concise form.

In Section 2, we discuss our testing procedure, and in Section 3 and 4, we compare the efficiency and to what extent tolerance proportionality holds. Then in Section 5, we compare the accuracy of the local error estimate and stepsize prediction. Following this, in Sections 6 and 7, we illustrate the effects of mild stiffness and of varying the scheme for selecting the stepsize. Finally, in Section 8 we summarize our conclusions and discuss their implications.

2. IMPLEMENTATION AND TESTING PROCEDURE

The pairs are implemented in a simple integrator consisting of a supervisor routine (RKDE) and a step integrator (RKSTP). On each accepted step, RKDE calculates the weights for the local error test and calls RKSTP. This routine attempts steps, until either the local error test is passed or the minimum stepsize is reached. On each attempted step the stages and the error estimate are calculated. If the step is accepted, x and **y** are updated and a new stepsize is selected according to the formula

$$\max\left\{h_{\min}, \min\left\{\alpha h_{\text{old}}, \beta(\text{TOL}/\text{est})^{1/p} h_{\text{old}}, h_{\max}\right\}\right\}$$

where $h_{\min}$ is the minimum stepsize allowed, $\alpha > 1$, $0 < \beta < 1$ is a safety factor, TOL is the local error tolerance, est is the weighted norm of the local error estimate, $h_{\max}$ is the maximum stepsize allowed, and h_{old} is the previous stepsize. The value of $h_{\min}$ is calculated on each step as in DVERK [10] (the formula for $h_{\min}$ is given in the appendix). The weighted norm of a vector is obtained by multiplying each component of the vector by the weight for the component and then calculating the norm of the resulting vector. The value of α is the maximum permissible ratio of consecutive attempted stepsizes. If the step is rejected, a new stepsize is calculated in one of two ways, depending on the number of consecutive rejected steps (denoted by n_r) at a point. If $n_r \leq m_{\text{opt}}$, where $m_{\text{opt}} \geq 0$, the new stepsize is selected as

$$\max\left\{j_{\text{old}}/\alpha, \beta(\text{TOL}/\text{est})^{1/p} h_{\text{old}}, h_{\min}\right\}.$$

For $n_r > m_{\text{opt}}$, the new stepsize is $\max\{h_{\min}, \alpha^{-1} h_{\text{old}}\}$. We refer to the parts of the above formulas that depend on the ratio TOL/est as the locally optimal formula. The type of weights and norm in the local error test and the values of α, β, $h_{\max}$, and m_{opt} are specified by the user. Unless stated otherwise, we use absolute weights (i.e., the weights in the weighted norm are all one), the maximum norm, $\alpha = 2$, $\beta = 0.9$, $h_{\max} = 20$, and $m_{\text{opt}} = 1$. Note that the stepsize is selected solely according to the above formulas. There is no explicit detection and handing of stiffness. The integrator is written so that the user can test a new pair merely by supplying the coefficients. Writing the integrator this way means the overhead is greater. But since we do not use the overhead to compare the pairs, our conclusions are unaffected.

The extensions to DETEST are of four types. We add an outer shell to enable us to test several pairs in one execution of the package. The output form DETEST is reformatted so it is easily used as input to programs that perform the comparisons. The number of rejected steps and more detailed information about the size of the error estimate relative to TOL is provided. Finally, we replace the (5, 6) pair of the integrator in DETEST that finds the true solution by the (7, 8) pair of Prince and Dormand. Preliminary testing showed the (7, 8) pair failed less often and gave more accurate global error estimates than the (5, 6) pair.

In our analysis, we use the normalized efficiency statistics for the number of derivative evaluations required to complete the integration. The statistics for a given problem are obtained as follows. It is assumed that the global error (either endpoint or maximum) satisfies the relationship

$$\text{global error} = C\,\text{TOL}^E$$

where the exponent E and the constant of proportionality C depend on the method and the problem. After the global error is found for several tolerances, values for E and C are found in a least squares sense. The values of C and E are then used to estimate the number of derivative evaluations required to achieve a prescribed global error. The number of derivative evaluations is found for a range of global errors to give the normalized efficiency statistics. Since the statistics are obtained from a least squares fit, it is important to know how reliable the fit is. This can be measured using the root mean square (rms) error of the fit, with a smaller value generally indicating the fit is more reliable.

To illustrate the behavior of the rms error, we solve problems A1, ..., A5, B1, ..., B5, D1, ..., D5, E1, ..., E5 (all scaled) of DETEST for tolerance ranges of $10^{-i/2}$, $i = 6, \ldots, 12$, $10^{i/2}$, $i = 12, \ldots, 20$, and 10^{-i}, $i = 3, \ldots, 10$. The C set of problems is omitted from these results because some problems in the set are mildly stiff (in this case we cannot expect tolerance proportionality because the stepsize is being limited by stability requirements and not accuracy requirements). The three tolerance ranges are used throughout the testing and are denoted by RL, RS, and RA, respectively. The R is a mnemonic for range while the L, S and A are mnemonics for lax, severe, and all tolerances.

The least squares fit assumes the principal term in the global error expansion dominates the higher order terms. The higher order terms will generally have less effect as the tolerance is made more severe and as the order is decreased (with the tolerance held constant), because the stepsize will generally be smaller in both cases. Hence we can expect the rms error to be smaller for RS than for RL and RA and the rms error to increase with the order. We find the rms error for RS is usually smaller than the rms error for RL and RA, and the rms error for N34 is generally smaller than for the other pairs. Also, for RS the rms error of the two (5, 6) pairs is generally smaller than for the (7, 8) pair. Each of these observations suggests the required behavior of the principal term and that our analysis will be more reliable for low orders and severe tolerances.

3. EFFICIENCY

To compare the efficiency of the pairs on nonstiff problems, we first solve problems A1, . . . , A5, B1, . . . , B5, D1, . . . , D5, E1, . . . , E5 of DETEST (all scaled) for RA. We refer to these problems as problems 1, . . . , 20. Next, we take the pairs two at a time and form the intersection of the values of the expected accuracy (from the normalized efficiency statistics) for each problem. For these values, we take the estimated number of derivative evaluations and divide the larger value by the smaller value. We then subtract one from this number to give the efficiency gain. If the second pair in the set uses more evaluations, we multiply the gain by negative one. Finally, we multiply the gains by a suitable scaling factor, round the result to the nearest integer, and enter the numbers on a graph of the expected accuracy against the problem number.

To illustrate the calculation of efficiency gains, suppose problem 10 is solved for RL using N34 and F45. The table of normalized efficiency statistics (formed from the endpoint global error) for N34 has entries for expected accuracies of 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, while the table for F45 has entries for 10^{-3} and 10^{-4}. The intersection of the values is therefore $\{10^{-3}, 10^{-4}\}$. For these values, N34 uses 352 and 529 evaluations, respectively, while F45 uses 274 and 433 evaluations. This gives efficiency gains of 0.28 and 0.22.

Since we are testing 6 pairs, 15 sets of efficiency gains are possible. This can be reduced to 5 if the second pair in each set is the same. If this pair is one of the (5, 6) pairs, we find that our conclusions about the relative efficiency of the pairs are similar to those obtained using all 15 sets.

Figures 1(a)-(e) contain the efficiency gains of the pairs relative to the Prince and Dormand (5, 6) pair for RA using the endpoint global error. In the gains, unity represents a 10-percent difference in efficiency. Since all the entries in Figure 1(a) are nonnegative, PD56 is at least as efficient as N34 at all accuracy requirements. With few exceptions, the same result holds for the (4, 5) pairs. However, for PD78 many of the entries at the lax accuracy requirements are zero or negative. Hence, PD78 is often more efficient than PD56 at lax accuracy requirements, although the gain is small. At severe accuracy requirements, we see that a higher order pair is generally more efficient, and it becomes increasingly more efficient as the accuracy requirement becomes more severe. For the (4, 5) and (5, 6) pairs, the Dormand and Prince pair at each order is generally more efficient. The results for PD78 at lax accuracy requirements are similar to those of Prince and Dormand [11]. They found PD78 was generally more efficient than the other pairs tested, including the (8, 9) pairs of Fehlberg [6] and Verner [13].

We initially attempted to find the efficiency gains using TOL = 10^{-i}, $i = 2, \ldots, 10$. However, the two (4, 5) pairs perform poorly for TOL = 10^{-2} on some of the orbit problems. On problem D1 with TOL = 10^{-2}, F45 uses 1000 steps without completing the integration. From approximately $x = 17$ onwards, the stepsize for accepted steps slowly approaches zero. A similar difficulty occurs on problems D2, D3, and D5. For TOL < 10^{-22}, all five orbit problems are integrated using considerably fewer steps, although the number

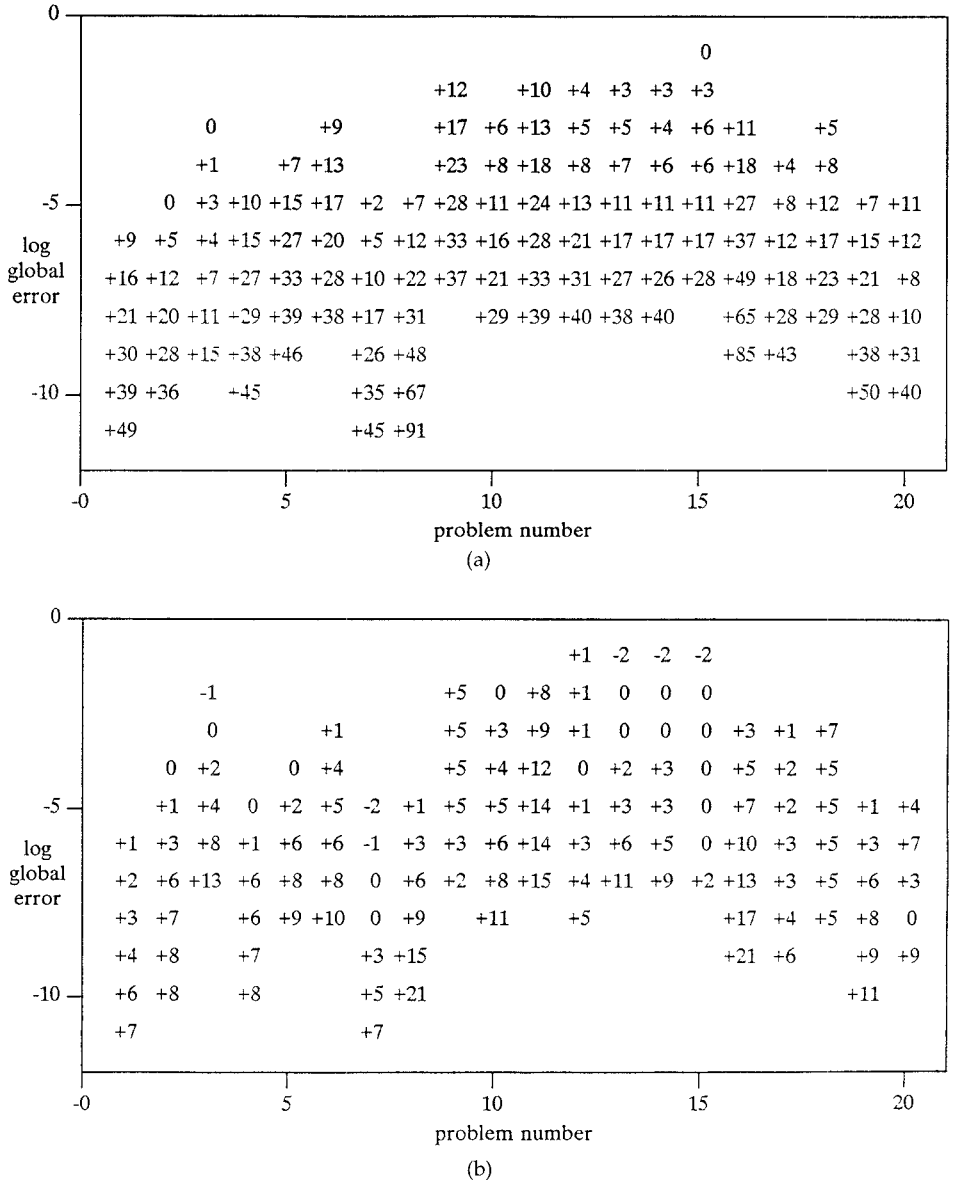
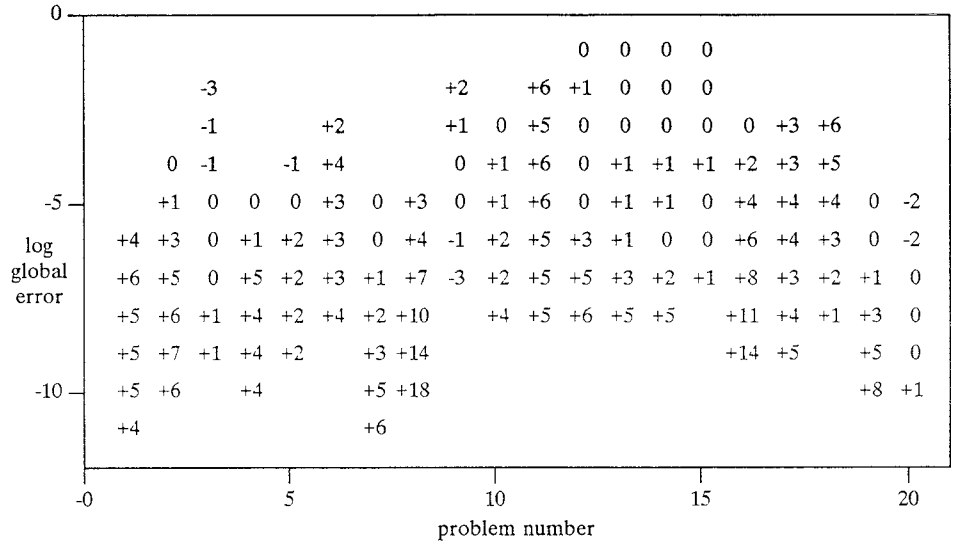
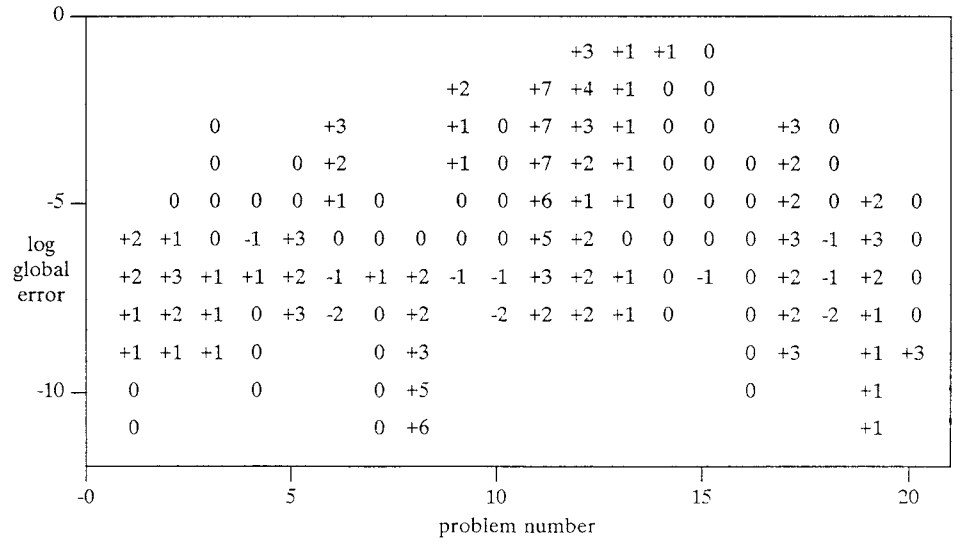


Fig. 1. (a) The efficiency gains for N34 relative to P56 on problems one to twenty. Unity represents ten percent. (b) The efficiency gains for F45 relative to PD56 on problems one to twenty. Unity represents ten percent. (c) The efficiency gains for DP45 relative to PD56 on problems one to twenty. Unity represents ten percent. (d) The efficiency gains for V56 relative to PD56 on problems one to twenty. Unity represents ten percent. (e) The efficiency gains for PD78 relative to PD56 for problems one to twenty. Unity represents ten percent.



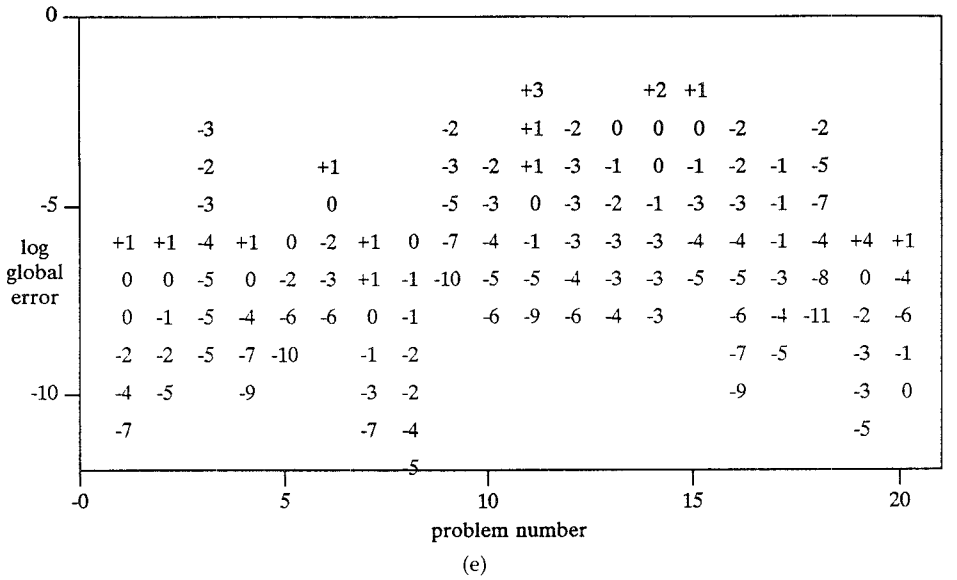
(c)



(d)

Fig. 1—*Continued*

of steps does not increase monotonically as the tolerance becomes more severe. The Dormand and Prince (4, 5) pair does not fare as badly. On D1 with TOL = 10^{-2} the pair takes 606 steps, while for TOL = $10^{-2\,2}$, 32 steps are taken. DP45 has little difficulty integrating the remaining D problems for TOL $\leq 10^{-2}$.



(e)

Fig. 1—*Continued*

4. TOLERANCE PROPORTIONALITY

One desirable property of an integrator is that the global error satisfy

$$\text{global error} = C\,\text{TOL}^E$$

where E is close to one (ideally equal to one) and C is a constant. This is commonly referred to as tolerance proportionality. In Figures 2(a)–(d) we give $|E-1|$ for problems $1,\ldots,20$ on RS, where E is that used in the normalized efficiency statistics. Figure 2(a) contains the values for N34 (solid line), F45 (dashed), and DP45 (dotted) using the endpoint global error. Figure 2(b) contains the values for V56 (solid line), PD56 (dashed), and PD78 (dotted) using the endpoint global error. Figures 2(c) and (d) contain the values for the maximum global error.

For N34, $E-1$ is significantly closer to zero than for the other pairs. We attribute this to the low order (hence smaller stepsize) of the pair and the fact that Nørsett (private communication) chose the coefficients of the pair so it had an accurate error estimate. The (5, 6) and (7, 8) pairs do not differ greatly on most problems. For all pairs there is generally little difference between the results for the endpoint and maximum global error, except for problem 1.

To quantify some of the differences in the values for E, we give

$$\frac{\Sigma|E-1|}{20}, \qquad \left[\frac{\Sigma(E-1)^2}{20}\right]^{-1/2}, \qquad \max\{|E-1|\}$$

in Table I.

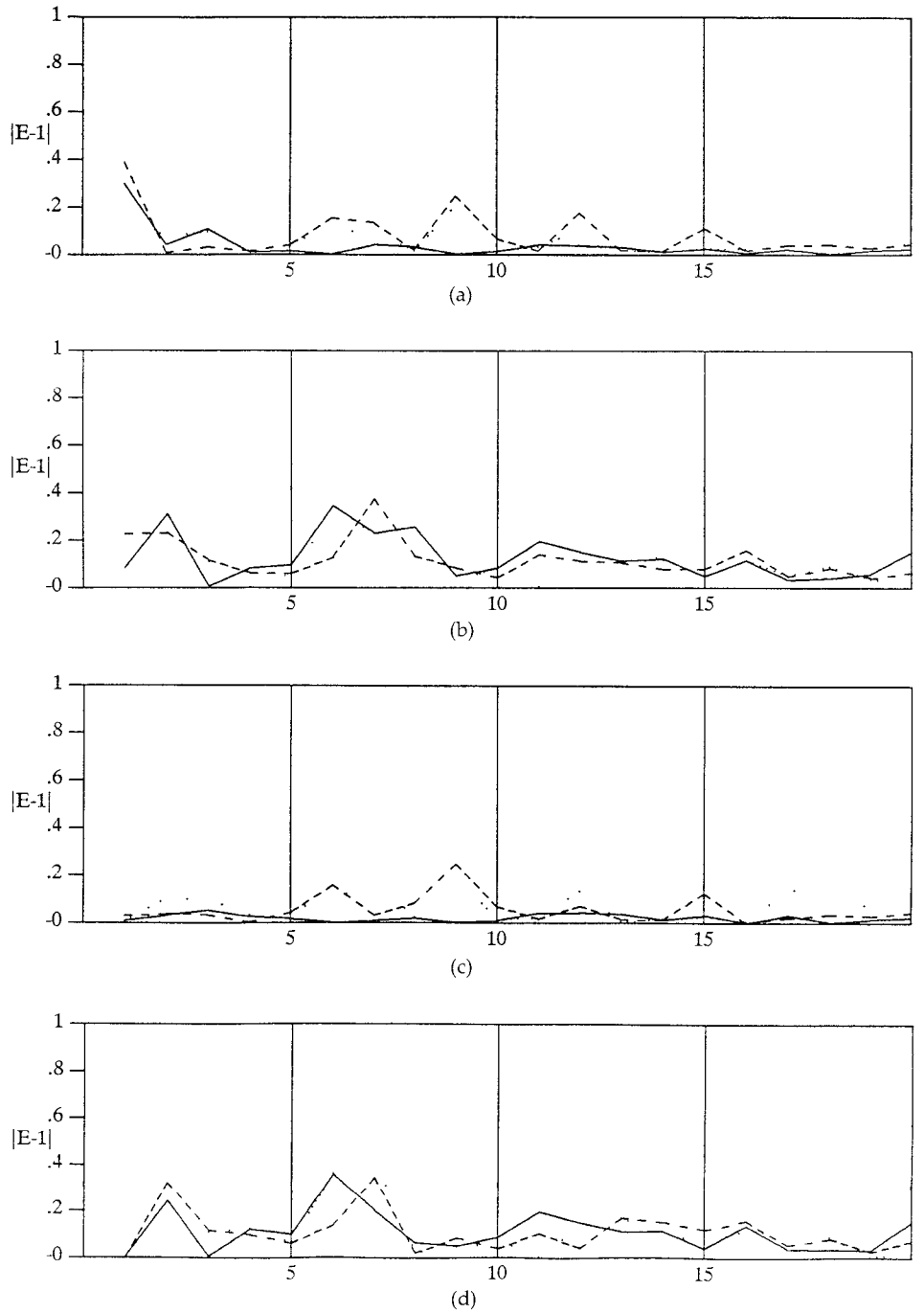


Fig. 2. $|E1|$ for problems 1,, . . , 20 on RS. (a) Endpoint error. Solid curve: N34; dashed curve: F45; dotted curve: DP45. (b) Endpoint error. Solid curve: V56; dashed curve: PD56; dotted curve: PD78. (c) and (d) Maximum error.

Table I. A Summary of $|E-1|$ for Problems One to Twenty on RS

Pair	Endpoint Error $\frac{\Sigma\|E-1\|}{20}$	Endpoint Error $\left[\frac{\Sigma(E-1)^2}{20}\right]^{1/2}$	Endpoint Error $\max\{\|E-1\|\}$	Maximum Error $\frac{\Sigma\|E-1\|}{20}$	Maximum Error $\left[\frac{\Sigma(E-1)^2}{20}\right]^{1/2}$	Maximum Error $\max\{\|E-1\|\}$
N34	0.0388	0.0749	0.2990	0.0200	0.0250	0.0520
F45	0.0800	0.1244	0.3880	0.0536	0.0793	0.2460
DP45	0.0784	0.0976	0.2070	0.0693	0.0912	0.1810
V56	0.1280	0.1579	0.3460	0.1107	0.1406	0.3570
PD56	0.1176	0.1419	0.3760	0.1085	0.1396	0.3440
PD78	0.1112	0.1418	0.3200	0.0987	0.1468	0.4230

Table II. Results for the Linear Regression of $|E-1|$ Against the rms Error for Problems One to Twenty on RS

Pair	Endpoint a_0	Endpoint a_1	Endpoint r	Maximum a_0	Maximum a_1	Maximum r
N34	0.145	0.054	0.575	0.061	0.018	0.788
F45	0.249	0.117	0.584	0.134	0.048	0.437
DP45	0.105	0.020	0.165	0.111	0.026	0.242
V56	0.161	0.023	0.135	0.188	0.049	0.285
PD56	0.259	0.112	0.654	0.238	0.088	0.611
PD78	0.163	0.058	0.310	0.208	0.105	0.437

The first two quantities measure the average deviation from tolerance proportionality while the third quantity measures the maximum deviation. The average deviations for N34 are significantly smaller than for the other pairs. However, because of its poor performance on problem 1 the maximum deviation for the endpoint error is greater than that of DP45. F45 has 4 deviations (in Table I) that are greater than those for DP45, while the average deviations for the (5, 6) and (7, 8) pairs are similar to one another.

We end this section by examining whether there is any correlation between $|E-1|$ and the rms error (from the normalized efficiency statistics), and whether larger values of the rms error are associated with larger values of $|E-1|$. Table II contains a_0 and a_1 in the linear regression $a_0 + a_1$ (rms error), along with the coefficient of correlation (r).

We see that $|E-1|$ depends weakly on the rms error, and that a_1 for the maximum global error increases with the order. The best correlation occurs for N34 and PD56, while for DP45 and V56 there is little correlation. For N34, the coefficient of correlation is smaller for the endpoint global error than the maximum global error primarily because of problem one.

5. LOCAL ERROR ESTIMATE

Irrespective of whether local extrapolation is used, most embedded pairs are derived and implemented so that the local error in the lower order formula is controlled directly. Hence, it is important to measure the accuracy of this

estimate. Related to this is how well the locally optimal formula (see Section 2) predicts the stepsize.

For an embedded pair using formulas of orders $p-1$ and p, the local error estimate for the lower order formula is

$$h^p \sum_k \hat{T}_{p,k}\mathbf{D}_{p,k} + h^{p+1} \sum_k (\hat{T}_{p+1,k} - T_{p+1,k})\mathbf{D}_{p+1,k} + O(h^{p+2})$$

and the true local error in the lower order formula is

$$h^p \sum_k \hat{T}_{p,k}\mathbf{D}_{p,k} + h^{p+1} \sum_k \hat{T}_{p+1,k}\mathbf{D}_{p+1,k} + O(h^{p+2})$$

where the $\hat{T}_{q,k}$ and $T_{q,k}$ ($q = p, p+1, \ldots$) are the truncation coefficients of order q, and the $\mathbf{D}_{q,k}$ are the corresponding elementary differentials.

If the $|T_{q,k}|$ are not sufficiently small relative to the $|T_{q,k}|$, especially for $q = p+1$, the estimated and true local errors can differ significantly, particularly for large stepsizes. This can lead to erratic behavior of the global error as a function of TOL, making it difficult to estimate the global error reliably using tolerance proportionally. The need for an accurate error estimate is recognized by Verner [13] and Prince and Dormand [2] as an important criteria to satisfy when selecting the values of the free parameters.

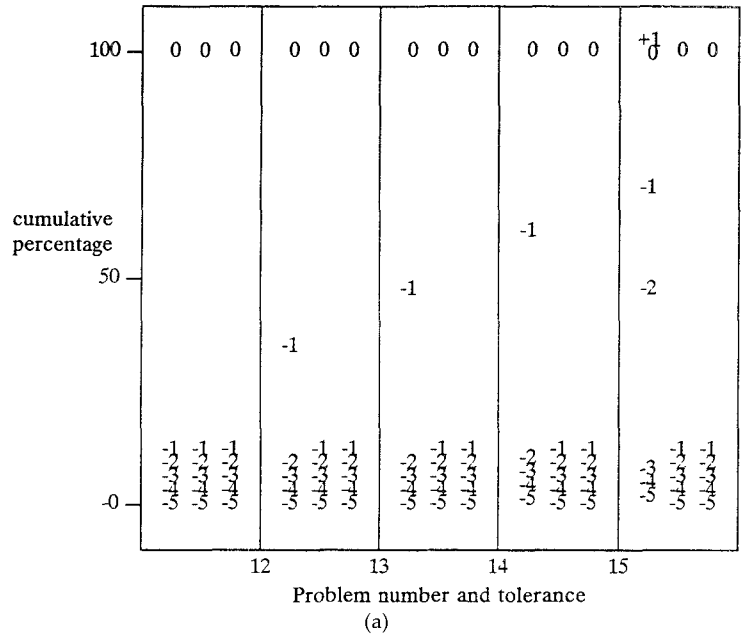
To test the accuracy of the local error estimate, we solve problems $11, \ldots, 15$ (the orbit problems) using TOL = 10^{-3}, 10^{-6}, 10^{-9}. These problems are chosen because as a group they provide a severe test of the accuracy of the local error estimate (see Prince and Dormand [5], for example). For each problem and tolerance, we count the number of accepted steps for which

$$\frac{\text{LE}}{\text{TOL}} \le 2^{-5}, \qquad 2^{j-1} < \frac{\text{LE}}{\text{TOL}} \le 2^j, \quad j = -4, \ldots, 5, \qquad 2^5 < \frac{\text{LE}}{\text{TOL}}$$

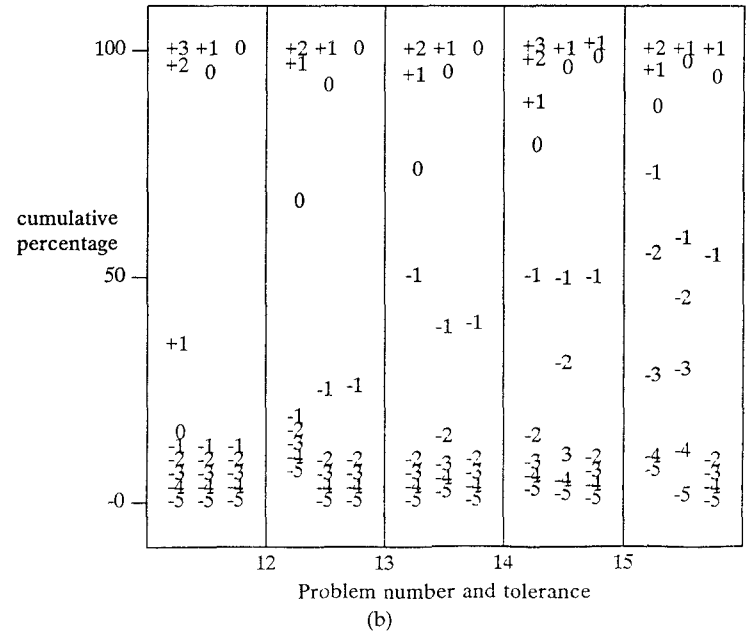
where LE is the weighted norm of the true local error in the lower order formula. This data can be arranged in a histogram of 12 intervals. If LE is generally close to (but less than) TOL and $\beta = 1$, the histogram will have a narrow peak centered on the sixth interval ($[1/2, 1]$). If β is 0.9, as we have in most of our tests, the peak will shift to the left, and the shift will increase with the order. But even for the (7, 8) pair, the shift will be only one interval since 0.9^8 is approximately $1/2$. If LE is often significantly smaller or larger than TOL, the peak will be less prominent.

In Figures 3(a)–(f) we give the cumulative percentage (CP) for each histogram. To form this, we calculate the number of steps with LE/TOL $\le 2^j$, $j = -5, \ldots, 5$ (if the twelfth interval is not empty, the CP is taken as 100). Then we place the value of j on the graph of the CP against the problem number and tolerance as follows. If an interval is empty, the value of j is placed immediately above the previous one. If two values in a column overlap, we move the one with the higher value of CP immediately above the other. These two conventions mean some of the points are not correctly placed on the graphs. This discrepancy has little effect on our conclusions.

To illustrate how the graphs are formed, consider solving N34 on problem 11 for TOL = 10^{-3} using N34 (the first column in Figure 3(a)). The number

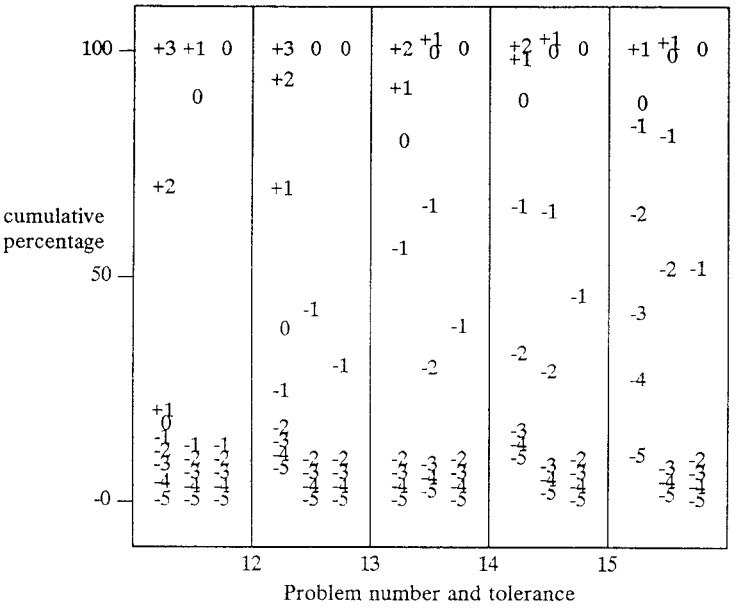


(a)

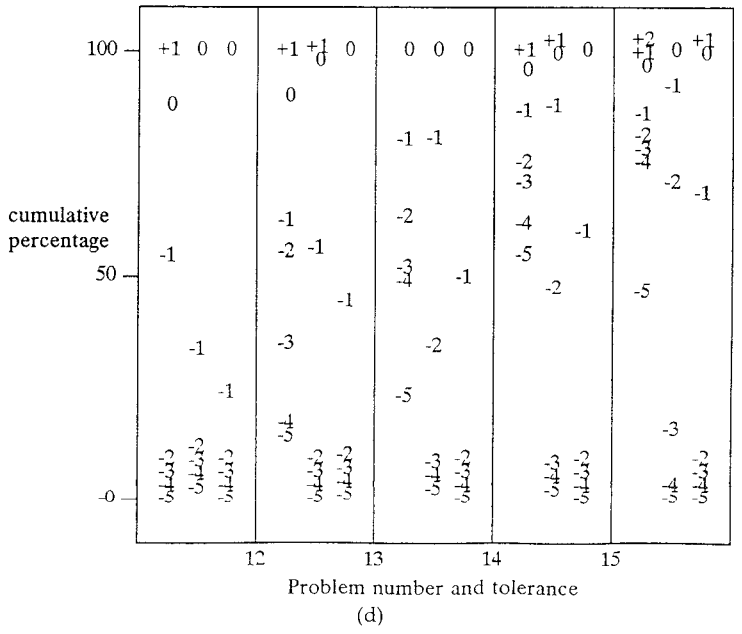


(b)

Fig. 3. (a), (b) A summary of LE/TOL for N34 and F45 applied to problems eleven to fifteen with TOL = 10^{-3}, 10^{-6}, and 10^{-9}. See the text for an explanation of the entries. (c), (d) A summary of LE/TOL for DP45 and V56 applied to problems eleven to fifteen with TOL = 10^{-3}, 10^{-6}, 10^{-9}. See the text for an explanation of the entries. (e), (f) A summary of LE/TOL for PD56 and PD78 pairs applied to problems eleven to fifteen with TOL = 10^{-3}, 10^{-6}, 10^{-9}. See the text for an explanation of the entries.

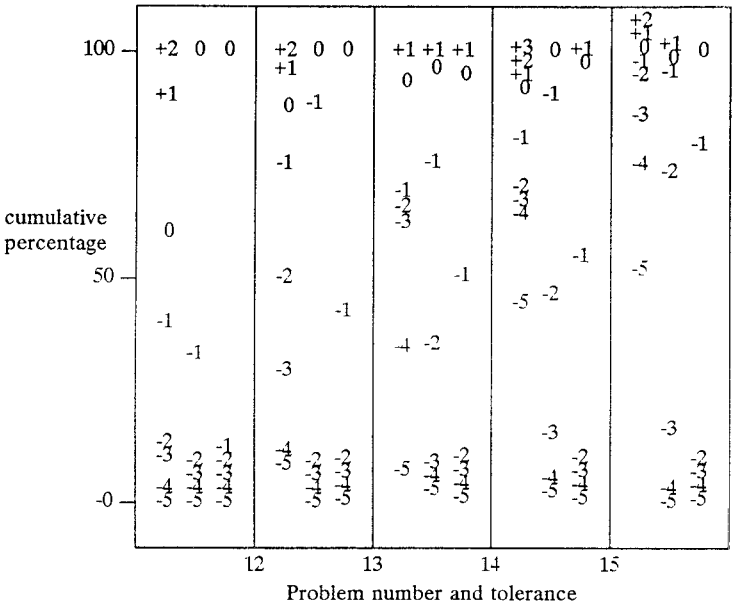


(c)

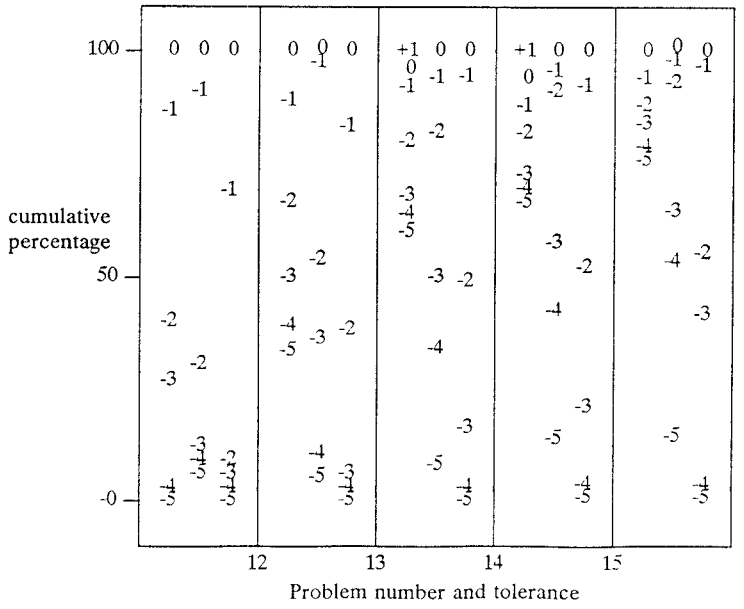


(d)

Fig 3—*Continued*



(e)



(f)

Fig. 3—*Continued*

of steps in the 12 intervals of the histogram are 0, 0, 0, 2, 5, 57, 0, 0, 0, 0, 0, 0, respectively. The CP for the first 3 intervals is 0 and we place the numbers -5, -4, -3 immediately above one another. The CP for the fourth and fifth intervals is 3.1 and 10.9, respectively. If -2, -1 are placed correctly, they overlap with -5, -4, and -3. Instead, we place -2 and -1 immediately above the -3. The CP for the sixth interval is 100 and we place 0 at the top of the graph.

With the data graphed in this way, several properties of the local error estimate are readily discerned. If LE/TOL is generally close to (but less than) 1, the 0 entries will occur near the top of the graph. N34 achieves this extremely well with all but one 0 occurring at 100 percent. Then in order of decreasing accuracy, we have PD78, V56, DP45, and PD56 of similar accuracy, and finally F45. We also see fewer positive integers occur at 100 percent as the tolerance decreases. This is because the principal term in the error estimate is more dominant, which makes the estimate more accurate.

If LE/TOL is less but not significantly less than 1, the negative entries will occur near the bottom, although -1 and possibly -2 for the higher order pairs are not required to. Once again N34 achieves this well. We also see two interesting patterns with the -1's for N34. As the eccentricity of the orbits increases (i.e., going from problem 11 to 15) or the tolerance increases, the -1's move up the graph. One reason for the former is that as the eccentricity increases the fraction of rejected steps increases (see below). Hence, because of the conservative nature of the stepsize selection, the stepsize is smaller than it need be. The latter effect follows from the dominance of the principal term in the error estimate. For the other pairs, the -1's move up the graphs as the order increases. This is partly due to β being smaller than 1. Finally, the other negative values show the same general dependence on the eccentricity as -1.

As well as having an accurate error estimate, we would like the locally optimal formula to predict stepsizes that lead to few rejected steps to avoid possible inefficiencies. Table III below contains a summary of the number of rejected steps for problems 11 to 15 with TOL = 10^{-3}, 10^{-6}, 10^{-9}. For each pair, problem, and tolerance, we give the number of times exactly one and two consecutive rejected steps occur at the same point (i.e., $n_r = 1$ and 2), as a percentage of the number of steps.

As might be anticipated from its low order and the results obtained so far, N34 generally has the smallest fraction of rejected steps. There are no rejected steps for TOL = 10^{-6}, 10^{-9}, and n_r is never 2. The (4, 5) pairs have no rejected steps for TOL = 10^{-9}. But for TOL = 10^{-3} on problems 14 and 15, both pairs have steps with $n_r = 2$. Otherwise, F45 generally has a smaller fraction of rejected steps than DP45. The remaining pairs have rejected steps at all tolerances. If we count $n_r = 2$ as two rejected steps, V56 generally has a smaller fraction of rejected steps than PD56. This is especially true for TOL = 10^{-9}. PD78 generally has a greater fraction of rejected steps than the two (5, 6) pairs.

Clearly the efficiency of the pairs can be improved, especially for the high

Table III. The Number of Steps with $n_r = 1, 2$ as a Percentage of the Number of Accepted Steps for Problems Sixteen to Twenty

Pair	TOL	11		12		13		14		15	
N34	10^{-3}	0	0	28	0	32	0	31	0	34	0
	10^{-6}	0	0	0	0	0	0	0	0	0	0
	10^{-9}	0	0	0	0	0	0	0	0	0	0
F45	10^{-3}	11	0	26	0	35	0	21	7	28	10
	10^{-6}	0	0	8	0	21	0	28	0	32	0
	10^{-9}	0	0	0	0	0	0	0	0	0	0
DP45	10^{-3}	34	0	27	0	35	0	32	4	22	11
	10^{-6}	0	0	14	0	24	0	30	0	34	0
	10^{-9}	0	0	0	0	0	0	0	0	0	0
V56	10^{-3}	8	0	20	6	34	2	29	9	21	12
	10^{-6}	0	0	19	0	30	0	33	0	35	0
	10^{-9}	0	0	0	0	0	0	7	0	16	0
PD56	10^{-3}	20	0	37	0	31	6	25	13	14	25
	10^{-6}	0	0	26	0	28	0	36	0	37	0
	10^{-9}	0	0	0	0	2	0	15	0	28	0
PD78	10^{-3}	20	0	33	0	28	12	21	15	16	22
	10^{-6}	27	0	33	0	40	0	37	0	39	0
	10^{-9}	0	0	23	0	34	0	35	0	36	0

order pairs, if the number of rejected steps can be decreased. We discuss this in Section 7.

6. STABILITY

On problem C1 of DETEST at lax tolerances, if the stepsize is selected using the formulas of Section 2, the stepsize will be limited by stability requirements. This means a similar number of steps is used for all lax tolerances. For more severe tolerances the accuracy requirement controls the stepsize, and the number of integration steps increases as the tolerance decreases. Problems of this type are referred to as mildly stiff.

When the stepsize is limited by stability requirements, the efficiency of pairs can be ranked by the size of their scaled stability region. This is the size of the stability region divided by the number of stages. In general, the larger the scaled region the more efficient the pair.

The size of a stability region can be measured in several ways. These include the magnitude of the intercepts of the stability boundary with the negative real and imaginary axes. The former is frequently used. The second and third columns of Table IV below contain the size of the unscaled and scaled stability region along the negative real axis. The scaled stability size for DP45 is one-sixth and not one-seventh of the unscaled size because the pair uses FSAL.

If the pairs are ranked using their scaled stability region, F45 should be the most efficient. Then in order of decreasing efficiency, we have N34 and DP45 of similar performance, V56 and PD56 of similar performance, and

Table IV. The Size of the Unscaled and Scaled Stability Region along the Negative Real Axis, as well as the Spectral Radius (μ) of the Equilibrium Matrix

Pair	Unscaled	Scaled	μ
N34	2.8	0.56	1.02
F45	3 7	0.62	0.99
DP45	3.3	0.55	1.02
V56	3.9	0.49	1.11
PD56	4.0	0.50	1 09
PD78	5.2	0 40	1.06

finally PD78. To test the accuracy of this ranking, problem C3 is solved using TOL = 10^{-i}, $i = 2, \ldots, 10$. The results for the endpoint global error are given below in Figure 4(a). For global errors greater than 10^{-3}, the problem is mildly stiff for all pairs, and the pairs rank in efficiency as stated above. For global errors less than 10^{-3}, the stepsize for N34 is not limited by the stability requirement and the number of evaluations increases as the tolerance decreases. A similar dependence starts for the remaining pairs at decreasingly smaller global errors as the order increases. It is not until the global error is approximately 10^{-8} that the (7, 8) pair is the most efficient. This last result is similar to that obtained by Dormand and Prince [2]. The final observation we make is for the (5, 6) pairs. The size of the stability region for the (5, 6) pairs differ little, and in Figure 4(a) the curves for the pairs are nearly coincident.

The behavior of a pair on a mildly stiff problem is not always as simple as in Figure 4(a). Figure 4(b) below contains the results for problem C4. A noticeable transition occurs between the region of mild stiffness and the asymptotic region. In the transition region, the global error changes little but the number of evaluations increases significantly.

For mildly stiff problems, explicit Runge–Kutta pairs can also be compared using the equilibrium theory developed by Higham and Hall (see [9], for example). They observed that the sequence of stepsizes can be one of two types. Either the stepsizes remain close to the boundary of the stability region, or they fluctuate about the boundary in a erratic manner. In the latter case, a significant fraction of the steps can be rejected, making the pair less efficient. For problems where the dominant eigenvalue of the Jacobian is real, the type of behavior which arises is independent of TOL and the type of norm used for the local error estimate. The desirable smooth stepsize sequence occurs if and only if $\mu < 1$, where μ is the spectral radius of the corresponding equilibrium matrix. This condition can be tested a priori and gives an additional criterion for comparing Runge–Kutta. An ideal pair will have $\mu < 1$ and possess a large stability radius. In the last column of Table III we give μ. Only F45 has $\mu < 1$, but the value of μ for the other pairs is not significantly greater than 1, and all pairs should have few rejected steps. This behavior is observed.

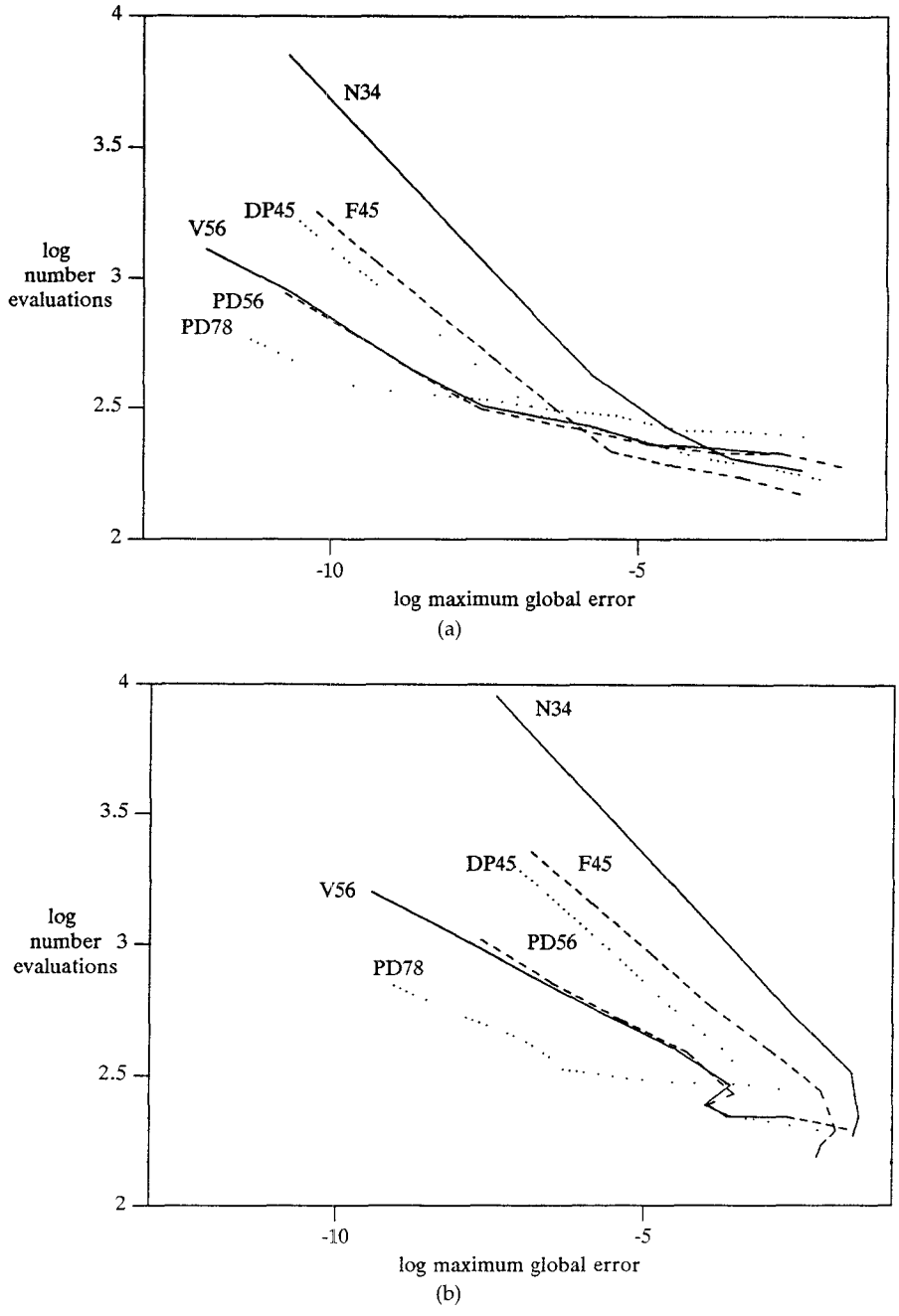


Fig. 4. (a), (b) The number of derivative evaluations against the endpoint global error on problem C3 (top) and C4 (bottom).

7. VARYING THE STEPSIZE SELECTION

In the previous sections we used the default values for β and m_{opt}. We now summarize the effect of changing β and m_{opt}.

All pairs have a significant fraction of rejected steps on the orbit problems, particularly on D5. A possible explanation for this is that the solution of the problems changes rapidly at some points and the locally optimal formula is unable to accurately predict stepsizes. One way to reduce the number of rejected steps is to make the stepsize prediction more conservative by reducing β. To test the effect of this, we reduce β to 0.7 and solve D5 on RA. With $\beta = 0.9$, N34 has rejected steps for only TOL = 10^{-3}, while F45 and DP45 have rejected steps for TOL = 10^{-i}, $i = 3, 4, 5, 6$, and V56, DP56, and PD78 have rejected steps for all of RA. When we reduce β to 0.7, all pairs have fewer rejected steps. If we use the number of function evaluations for a given global error as the measure of efficiency, the gain in efficiency with $\beta = 0.7$ ranges from approximately 5 to 30 percent.

Our default value of m_{opt} is 1, which is the same as in DVERK (Hull et al. [10]). This means that on the first rejected step the locally optimal formula is used, and on subsequent rejections (at the same point) the stepsize is reduced by a factor of α^{-1}. Other values for m_{opt} are also commonly used. For example, in DESTEP (Shampine and Gordon [12]) the stepsize is halved after the first rejected step ($m_{\text{opt}} = 0$). To illustrate some of the effects of having $m_{\text{opt}} \neq 1$, we solve the orbit problems with $m_{\text{opt}} = 0$ and 2 for tolerances of TOL = 10^{-3}, 10^{-6}, and 10^{-9}.

Decreasing m_{opt} from 1 to 0 decreases the fraction of steps with $n_r = 2$. But the fraction of steps with $n_r = 1$ increases for all pairs at TOL = 10^{-3}, 10^{-6} and for PD78 at TOL = 10^{-9}. The first result follows from the more conservative nature of the stepsize selection. The reason for the second result is not as obvious. When a step is accepted after a rejected step, the predicted stepsize for the next step is often too large. This leads to sequences of alternating rejected and accepted steps which increase the fraction of steps with $n_r = 1$. Increasing m_{opt} from one to two decreases the fraction of steps with $n_r = 1$, but increases the fraction with $n_r = 2$.

8. DISCUSSION

We performed numerical testing of six explicit Runge–Kutta pairs ranging in order from a (3, 4) pair to a (7, 8) pair. All the test problems had smooth solutions and we assumed dense output was not required. The pairs were implemented in a uniform way. In particular, the stepsize selection for all pairs was based on the locally optimal formula. We tested the efficiency of the pairs, to what extent tolerance proportionality held, the accuracy of the local error estimate and stepsize prediction, and the performance on mildly stiff problems. We also showed, for these pairs, how the performance could be altered noticeably by making simple changes to the stepsize selection strategy. As part of the work, we demonstrated new ways of presenting numerical comparisons.

Although we tested individual pairs of a class and not classes, we think some general conclusions can be made about the classes. The first concerns the efficiency of the pairs on truly nonstiff problems for which the solution does not change rapidly. The (7, 8) pair of Prince and Dormand generally uses fewer derivative evaluations than the (5, 6) pairs. Since the work of Verner [13] suggests existing (5, 6) pairs are close to the most efficient possible, we think (7, 8) pairs will generally be more efficient than (5, 6) pairs, provided the coefficients are chosen suitably.

We also found that the (7, 8) pair was more efficient than the lower order pairs on problems whose solution changes rapidly (the orbit problems). However, this conclusion may not hold in general, since our results suggest the percentage of rejected steps will usually increase with the order. In this case, the efficiency of a high order pair relative to a low order one will be reduced, and it may be possible on some problems for the low order pair to be more efficient. We intend investigating ways of ensuring that high order pairs retain most or all of their efficiency. We anticipate this will involve modifying the scheme for selecting the stepsize, as well as the criteria used to select the coefficients of a pair.

A (7, 8) pair may also not be more efficient than a lower order pair when dense output is required. The output can be obtained by either hitting the output point exactly or by using interpolants. In the former case, the stepsize will be the same for all pairs. Hence, the cost of advancing a step using a (7, 8) pair will be greater than that for a (5, 6) pair, and the (7, 8) pair will generally be less efficient. If interpolants are used, the (5, 6) and (7, 8) pairs will require evaluations in addition to those of the underlying pair. If the (7, 8) pair requires more such evaluations than the (5, 6) pair, the (7, 8) pair will be less efficient relative to the (5, 6) pair. We are investigating this issue.

Our testing showed that the (3, 4) pair of Nørsett generally has a more accurate local error estimate at lax tolerances than the other pairs and displays better tolerance proportionality. Hence the (3, 4) pair would be useful at lax tolerances on problems for which accurate error estimates were required.

On mildly stiff problems with real eigenvalues, high order pairs are less efficient than low order pairs at lax tolerances. Dormand and Prince [2] showed their (4, 5) and (5, 6) pairs could be made more efficient by extending the stability region along the negative real axis. Although the efficiency is improved, this approach has the following disadvantage. Choosing the coefficients of the pair so that the stability region is extended will generally make the pair less efficient on nonstiff problems (see the numerical results of Dormand and Prince [2], for example). We think mildly stiff problems can be handled more effectively by detecting the stiffness and switching to a low order pair.

APPENDIX

We give the tableau of the Fehlberg (4, 5) pair and the formula for the minimum stepsize used in the testing (see Figure 5).

Figure 5

$\frac{0}{1}$						
$\frac{1}{4}$	$\frac{1}{4}$					
$\frac{3}{8}$	$\frac{3}{32}$	$\frac{9}{32}$				
$\frac{12}{13}$	$\frac{1932}{2197}$	$\frac{-7200}{2197}$	$\frac{7296}{2197}$			
1	$\frac{439}{216}$	-8	$\frac{3680}{513}$	$\frac{-845}{4104}$		
$\frac{1}{2}$	$\frac{-8}{27}$	2	$\frac{-3544}{2565}$	$\frac{1859}{4104}$	$\frac{-11}{40}$	
4^{th}	$\frac{25}{216}$	0	$\frac{1408}{2565}$	$\frac{2197}{4104}$	$\frac{-1}{5}$	0
5^{th}	$\frac{16}{135}$	0	$\frac{6656}{12825}$	$\frac{28561}{56430}$	$\frac{-9}{50}$	$\frac{2}{55}$

The minimum stepsize ($h_{\min}$) is intended to prevent the integrator from using a stepsize for which the roundoff errors are significant. We calculate $h_{\min}$ as

$$C_1 \max\{C_2, u \max\{|\, y_i \,|/\mathrm{TOL}, |\, x_i \,|\}\}$$

where C_1, C_2 are constants that do not depend on the pair and problem, u is the unit roundoff, and $|\, y_i \,|$ is the weighted norm of y_i (the weights are the same as in the local error test.) The constant C_1 is a safety factor and C_2 is a very small positive machine number. Throughout the testing we use $C_1 = 10$ and $C_2 = 10^{-20}$

ACKNOWLEDGMENT

The author thanks W. Enright, K. Jackson, C. Christara, D. Higham, and R. Enenkel whose suggestions improved the first draft, and one of the referees who made a large number of detailed suggestions.

REFERENCES

1. Dormand, J. R., and Prince, P. J. A family of embedded Runge-Kutta formulae. *J. Comput Appl. Math.* **6** (1980), 19-26.
2. Dormand, J. R., and Prince, P. J. A reconsideration of some embedded Runge-Kutta formulae. *J. Comput. Appl. Math.* **15** (1986), 203-211.
3. Enright, W. H., and Hull, T. E. Test results on initial value methods for non-stiff ordinary differential equations. *SINUM* **13** (1976), 944-961.
4 Enright, W. H , Jackson, K. R., NØrsett, S. P., and Thomsen, P. G. Interpolants for Runge-Kutta formulas. *ACM Trans. Math. Softw.* **12** (1986), 193-218
5. Enright, W H., and Pryce, J. D. Two FORTRAN packages for assessing initial value methods *ACM Trans. Math. Softw.* **13** (1987), 1-27.
6. Fehlberg, E. Classical fifth, sixth, seventh, and eighth order Runge-Kutta formulas with stepsize control. Tech. Rep. TR R-287, NASA, 1968.

7. FEHLBERG, E. Klassische Runge-Kutta-Formeln funfter und siebenter Ordnung mit Schrittweiten-Kontrolle. *Computing* **4** (1969), 93-106.
8. FEHLBERG, E. Low order classical Runge-Kutta formulae with stepsize control and their application to some heat transfer problems. Tech. Rep. R-315, NASA, 1969.
9. HIGHAM, D. J., AND HALL, G. Runge-Kutta equilibrium theory for a mixed absolute relative error measure. Tech. Rep. TR 218/89, Department of Computer Science, University of Toronto, 1989.
10. HULL, T. E., ENRIGHT, W. H., AND JACKSON, K. R. User's guide for DVERK—A subroutine for solving nonstiff ODE's. Tech. Rep. TR 100/76, Department of Computer Science, University of Toronto, Canada, 1976.
11. PRINCE, P. J., AND DORMAND, J. R. High order embedded Runge-Kutta formulae. *J. Comput. Appl. Math.* **7** (1981), 67-76.
12. SHAMPINE, L. F., AND GORDON, M. K. Computer solution of ordinary differential equations: The initial value problem. W. H. Freeman, San Francisco, 1975, pp. 186-209.
13. VERNER, J. H. Explicit Runge-Kutta methods with estimates of the local truncation error. *SINUM* **15** (1978), 772-790.
14. VERNER, J. H. A contrast of some Runge-Kutta formula pairs. To appear in *SINUM*.

Received August 1988; accepted May 1990