IMA Journal of Numerical Analysis (1991) **11**, 457–480

Global Error versus Tolerance for Explicit Runge–Kutta Methods

DESMOND J. HIGHAM

Department of Computer Science, University of Toronto, Canada M5S 1A4

Dedicated to Professor A. R. Mitchell on the occasion of his 70th birthday

[Received 3 August 1990 and in revised form 6 December 1990]

Initial value solvers typically input a problem specification and an error tolerance, and output an approximate solution. Faced with this situation many users assume, or hope for, a linear relationship between the global error and the tolerance. In this paper we examine the potential for such 'tolerance proportionality' in existing explicit Runge–Kutta algorithms. We take account of recent developments in the derivation of high-order formulae, defect control strategies, and interpolants for continuous solution and first derivative approximations. Numerical examples are used to verify the theoretical predictions. The analysis draws on the work of Stetter, and the numerical testing makes use of the nonstiff DETEST package of Enright and Pryce.

1. Introduction

WE are concerned with the numerical solution of the nonstiff initial value problem

$$y'(t) = f(t, y(t)), \qquad y(t_0) = y_0 \in \mathbb{R}^N, \quad t_0 \leq t \leq t_{\text{end}}. \tag{1.1}$$

We assume that discrete numerical approximations are produced by a single 'one-pass' integration with an s-stage explicit Runge–Kutta (RK) formula. A typical step with such a formula advances the approximation $y_{n-1} \approx y(t_{n-1})$ to $y_n \approx y(t_n)$ according to

$$\begin{aligned} k_1 &= f(t_{n-1}, y_{n-1}), \\ k_i &= f\left(t_{n-1} + c_i h_n, y_{n-1} + h_n \sum_{j=1}^{i-1} a_{ij} k_j\right), \quad i = 2, \ldots, s, \\ y_n &= y_{n-1} + h_n \sum_{i=1}^{s} b_i k_i, \end{aligned} \tag{1.2}$$

where $h_n := t_n - t_{n-1}$ is the local stepsize, and the fixed scalars $\{a_{ij}, b_i, c_i\}_{i,j=1}^{s}$ are the coefficients of the RK formula. In its simplest form, an 'implementation' of the RK formula is a method for choosing the stepsizes h_n, and hence determining the mesh $\{t_n\}$. This choice involves a trade-off between accuracy and computational expense. Smaller stepsizes give greater accuracy, but larger stepsizes allow the integration to be completed with fewer steps. For this reason, modern codes

Current address: Department of Mathematics and Computer Science, University of Dundee, Dundee DD1 4HN.

allow the user to supply a tolerance parameter, δ, which gives an indication of the level of accuracy required. The codes then attempt to deliver a sufficiently accurate solution as efficiently as possible.

The fundamental measure of the error at each meshpoint t_n is the *global error* $y_n - y(t_n)$. In general, it is not possible to keep the global error within a prescribed bound in a single integration. (Once we have strayed from the true solution we begin to track a neighbouring, and possibly divergent solution curve.) Hence, codes normally control a locally based measure of the error, and so control the global error indirectly, in a problem-dependent manner. Perhaps the best that we can ask in this situation is that if a fixed problem (1.1) is solved repeatedly over a 'reasonable' range of tolerance values, then the global error should decrease linearly with δ. Such behaviour is, of course, exactly what an optimistic user would automatically expect. Following Stetter (1978, 1980) we refer to this characteristic as tolerance proportionality (TP). The aim of this work is to investigate the extent to which TP is likely to hold in modern RK codes. We pay special attention to high-order formulas, defect control techniques, and continuous extensions (interpolants). In the remainder of this section we outline the error control techniques that are in current use.

On a particular step from t_{n-1} to t_n, the RK formula inherits the approximation y_{n-1} as a starting value. Hence the formula can be regarded as following the *local solution* for the step $z_n(t)$, which is defined by

$$z_n'(t) = f(t, z_n(t)), \qquad z_n(t_{n-1}) = y_{n-1}.$$

It is therefore useful to define the *local error* for the step,

$$\mathrm{le}_n := y_n - z_n(t_n).$$

We say that a formula has order p if p is the largest integer such that the local error satisfies $\mathrm{le}_n = O(h_n^{p+1})$. Throughout this work we will suppose that the numerical solution is advanced with a pth order formula. (We also assume without further comment that the problem (1.1) is sufficiently smooth.) A result that will be needed later is that the local error in a pth order formula has an expansion of the form

$$\mathrm{le}_n = h_n^{p+1}\psi(y_{n-1}, t_{n-1}) + O(h_n^{p+2}), \tag{1.3}$$

where ψ is a smooth function.

Most of the error control techniques in existing codes attempt to ensure that some measure of the local error is small on each step. If $\mathrm{est}_n \leq \delta$, where est_n is a measure of a local error estimate, then the step is accepted, otherwise the step is repeated with a smaller stepsize h_n. The error estimate can be obtained by advancing from y_{n-1} to $\bar{y}_n$ with a subsidiary RK formula of a different order. The quantity $\|y_n - \bar{y}_n\|$ then gives an asymptotically correct approximation to the norm of the local error in the lower-order formula. Some codes use $\mathrm{est}_n = \|y_n - \bar{y}_n\|$, while others use $\mathrm{est}_n = \|y_n - \bar{y}_n\|/h_n$. The former choice is called error-per-step control, the latter error-per-unit-step. When the order of the subsidiary formula is less than p, we are estimating the local error in $\bar{y}_n$ while advancing with the higher-order approximation y_n. This is normally referred to as

local extrapolation. In the other case, when the subsidiary formula is of order greater than p, we are estimating the local error in the actual numerical solution y_n. There are thus four possible modes of local error control, which, following Shampine (1977), can be abbreviated to

EPS: No local extrapolation, error-per-step control.
EPUS: No local extrapolation, error-per-unit-step control.
XEPS: Local extrapolation, error-per-step control.
XEPUS: Local extrapolation, error-per-unit-step control.

Each of the four modes has been used in practice. XEPS is arguably the most popular of the four, and is regarded as the most efficient (Shampine, 1977).

Usually the Euclidean or infinity norm is used in the computation of est_n and, in general, the local error estimate will be premultiplied by a weighting matrix $\text{diag}\{w_i^{-1}\}$ before the norm is taken. The extremes of $w_i \equiv 1$ and $w_i = \frac{1}{2}\{|y_{n-1}|_i + |y_n|_i\}$ correspond to uniform absolute and uniform relative weights respectively. More general weighting schemes involve componentwise relative and absolute weighting parameters RTOL_i and ATOL_i, which may be specified by the user. For example, the code DERKF (Shampine & Watts, 1979) uses

$$w_i = \text{RTOL}_i \tfrac{1}{2}\{|y_{n-1}|_i + |y_n|_i\} + \text{ATOL}_i.$$

A summary of the weighting types used in several popular codes is given in Higham & Hall (1989). It is clear that our concept of tolerance proportionality must include the assumption that the weighting parameters remain the same as δ varies.

In addition to the discrete numerical solution $\{y_n\}$, it is frequently useful to have available a continuous function $q(t)$ which approximates $y(t)$ over the range $[t_0, t_{\text{end}}]$. This can be done by augmenting the RK formula (1.2) with a 'continuous extension' $q(t)$ of the form

$$q(t_{n-1} + \tau h_n) = y_{n-1} + \tau h_n \sum_{i=1}^{\hat{s}} b_i(\tau) k_i, \quad \tau \in (0, 1], \tag{1.4}$$

where each $b_i(\tau)$ is a polynomial in τ (see, for example, Dormand & Prince, 1986; Enright *et al.*, 1986; Higham, 1989*b*; Shampine, 1985, 1986). (If $\hat{s} > s$ then extra stages must be added to the basic RK formula.) The local approximations can be joined together in a piecewise fashion to give a global approximation to $y(t)$. Continuous extensions usually satisfy the conditions $q(t_{n-1}) = y_{n-1}$ and $q(t_n) = y_n$, and hence are referred to as interpolants. Modern interpolants also match the derivative data at the meshpoints; that is,

$$q'(t_{n-1}) = f(t_{n-1}, y_{n-1}) \quad \text{and} \quad q'(t_n) = f(t_n, y_n),$$

in which case the corresponding global approximation is a C^1 function. To assess the accuracy of an interpolant, we say that $q(t)$ has *local order* l if l is the largest integer such that for any fixed $\tau \in [0, 1]$

$$q(t_{n-1} + \tau h_n) - z_n(t_{n-1} + \tau h_n) = O(h_n^{l}).$$

The literature includes examples where $l = p + 1$, so that the local errors in the formula and the extension have the same order, and others where $l = p$. We will use the phrases 'higher order' and 'lower order' to distinguish between these two classes. For the latter class, since the global error in y_n behaves like $h_{\max}^p$, where $h_{\max}$ is the maximum stepsize, the *global error* in the discrete and continuous formulae should be compatible. Some consequences of the choice between $l = p + 1$ and $l = p$ will be discussed in later sections.

Given the availability of C^1 approximations, Enright (1989*a*) suggested an alternative to the local error control schemes described above (see also Enright, 1989*b*; Higham, 1989*a*, 1989*b*, for further developments). The idea is to form an interpolant and control a sample of the quantity

$$\operatorname{def}(t) := q'(t) - f(t, q(t)),$$

which we call the *defect* in $q(t)$. If we have an interpolant of local order l, and sample at a point $t_{n-1} + \tau^* h_n$, for some fixed $\tau^* \in (0, 1)$, then the defect sample satisfies

$$\operatorname{def}(t_{n-1} + \tau^* h_n) = h_n^{l-1} \Phi(y_{n-1}, t_{n-1}) + O(h_n^l), \tag{1.5}$$

where Φ is a smooth function. For a general continuous extension, the shape of the defect depends upon the problem and hence $\max_{[0,1]} \|\operatorname{def}(t_{n-1} + \tau h_n)\|$ cannot be controlled by sampling at a single prechosen point. However, numerical tests (see, for example, Enright, 1989*a*) have shown that the maximum is very rarely more than ten times the sampled value, if the sample point is carefully chosen. Special classes of interpolants were derived in Higham (1989*a*, 1989*b*) for which one sample gives an asymptotically correct estimate of the maximum defect. To obtain such a property, one must pay a rather high cost per step in the formation of $q(t)$.

In the next section we give some theoretical results relating the error control technique to the global errors in the discrete solution. The results are based on those of Stetter (1978, 1980). Section 3 then summarizes some numerical testing of various error control strategies in conjunction with formula pairs of orders (3, 4), (4, 5), and (7, 8). For these tests we make use of the nonstiff DETEST package (Enright & Pryce, 1987). In Section 4 we look at the way that the global errors in the continuous extensions vary with the tolerance. Predictions based on an asymptotic analysis are tested numerically. We consider the global errors in approximations to the first derivative, both from the discrete data $\{f(t_n, y_n)\}$ and the continuous extensions, in Section 5. Again, we present numerical results to test the asymptotic theory. Finally, in Section 6 we summarize and discuss our findings.

2. Theoretical results

What local constraints on the stepsize are necessary/sufficient for the *discrete* numerical solution to exhibit tolerance proportionality? In this section we reorganize and slightly extend some results of Stetter (1978, 1980) that relate to this question. To be specific, Theorem 2.1 below is contained in, and Theorems 2.2 and 2.3 are extensions of, Theorem 1 of Stetter (1978). (We mention that the

$o(\delta)$ term appearing in the statement of Theorem 1 of Stetter (1978) should be an $o(h_n\delta)$ term.) Corollaries 2.1 and 2.2 were outlined informally in Stetter (1980). Some related work on equivalent conditions for an approximate ODE solution can be found in Stewart (1970).

Before beginning an analysis, we must be clear about what we mean by TP for a discrete method. Since the location of the meshpoints $\{t_n\}$ will generally depend upon the tolerance, it does not make sense to ask for the 'global errors at the meshpoints' to decrease linearly with the tolerance. Instead, we should ask for the existence of a continuous interpolant $\eta(t)$ which passes through the meshpoint data $\{t_n, y_n\}$ and satisfies $\eta(t) - y(t) \approx v(t)\delta$, where $v(t)$ is independent of the tolerance δ. Note that $\eta(t)$ need not be computable, since we are only concerned with the meshpoint approximation (in this section).

It is worth mentioning at this stage that the interpolant defined by

$$\eta_I(t) := z_n(t) + \frac{(t - t_{n-1})}{h_n} \mathrm{le}_n, \quad t \in (t_{n-1}, t_n], \tag{2.1}$$

plays an important role in this work. Note that $\eta_I(t_{n-1}) = z_n(t_{n-1}) = y_{n-1}$ and $\eta_I(t_n) = z_n(t_n) + \mathrm{le}_n = y_n$, so we do indeed have a continuous interpolant through the mesh data. Following Stetter (1979) we will refer to $\eta_I(t)$ as the *ideal* interpolant. (It is ideal in the sense that, asymptotically, its defect has the smallest possible value on every step.) Note, however, that the first derivative of $\eta_I(t)$ can be discontinuous at each meshpoint.

In the following analysis we suppose that for any tolerance value δ there exists a corresponding mesh $\{t_n\}$ for the RK solution. The proofs include the implicit assumption that the stepsizes tend to zero as $\delta \to 0$; that is, $\max_n h_n = o(1)$ as $\delta \to 0$. We restrict the phrase 'piecewise continuous' to mean continuous except possibly at the meshpoints $\{t_n\}$, and the phrase 'piecewise C^1' to mean continuous over the whole range $[t_0, t_{\mathrm{end}}]$ with a first derivative which is continuous except possibly at the meshpoints $\{t_n\}$.

THEOREM 2.1 *Given the initial value problem* $y'(t) - f(t, y(t)) = 0$, $y(t_0) = y_0$, *suppose* $\eta(t)$ *is piecewise* C^1 *and satisfies* $\eta(t_0) = y_0$. *Let* $\varepsilon(t) := \eta(t) - y(t)$ *denote the global error in* $\eta(t)$. *Then the conditions* A *and* B *below are equivalent*:

Condition A: $\varepsilon(t) = v(t)\delta + g(t)$, *where* $v(t)$ *is* C^1 *and independent of* δ, *and* $g(t)$ *is piecewise* C^1 *with zeroth and first derivatives of* $o(\delta)$.

Condition B: $\eta'(t) - f(t, \eta(t)) = \gamma(t)\delta + s(t)$, *where* $\gamma(t)$ *is continuous and independent of* δ, *and* $s(t)$ *is piecewise continuous and* $o(\delta)$.

Proof. We introduce a third condition, C, and then prove that A $\Rightarrow$ B, B $\Rightarrow$ C, and C $\Rightarrow$ A.

Condition C: $\varepsilon'(t) - f_y(t, y(t))\varepsilon(t) = \gamma(t)\delta + u(t)$, where $\gamma(t)$ is the function appearing in condition B, and $u(t)$ is piecewise continuous and $o(\delta) + O(\varepsilon(t)^2)$.

A $\Rightarrow$ B: We have

$$\begin{aligned}\eta'(t) - f(t, \eta(t)) &= y'(t) + \varepsilon'(t) - f(t, y(t) + \varepsilon(t)) \\ &= y'(t) + \varepsilon'(t) - f(t, y(t)) - f_y(t, y(t))\varepsilon(t) + w(t) \\ &= \varepsilon'(t) - f_y(t, y(t))\varepsilon(t) + w(t),\end{aligned}$$

where $w(t) = O(\varepsilon(t)^2)$, and hence, from A, $w(t) = o(\delta)$. Using A in this equation we obtain

$$\eta'(t) - f(t, \eta(t)) = \delta[v'(t) - f_y(t, y(t))v(t)] + g'(t) - f_y(t, y(t))g(t) + w(t),$$

which has the required form.

B $\Rightarrow$ C: Subtracting the original ODE, $y'(t) - f(t, y(t)) = 0$, from B gives

$$\eta'(t) - y'(t) - [f(t, \eta(t)) - f(t, y(t))] = \gamma(t)\delta + s(t).$$

Using a Taylor expansion of $f(t, \eta(t)) = f(t, y(t) + \varepsilon(t))$ this becomes

$$\varepsilon'(t) - f_y(t, y(t))\varepsilon(t) + \bar{w}(t) = \gamma(t)\delta + s(t),$$

where $\bar{w}(t)$ is piecewise continuous and $O(\varepsilon(t)^2)$.

C $\Rightarrow$ A: Let $v(t)$ denote the unique solution to the linear initial value problem

$$v'(t) - f_y(t, y(t))v(t) = \gamma(t), \qquad v(t_0) = 0.$$

Then, from C, $\varepsilon(t) - \delta v(t)$ satisfies

$$(\varepsilon(t) - \delta v(t))' - f_y(t, y(t))(\varepsilon(t) - \delta v(t)) = u(t), \qquad \varepsilon(t_0) - \delta v(t_0) = 0.$$

Standard theory (see, for example, Ascher, Mattheij & Russell, 1988: p. 86) shows that this initial value problem has a solution of the form

$$\varepsilon(t) - \delta v(t) = Y(t) \int_{t_0}^{t} Y^{-1}(\mu)u(\mu)\, d\mu,$$

where the fundamental solution matrix $Y(t)$ is defined by

$$Y'(t) = f_y(t, y(t))Y(t), \qquad Y(t_0) = I.$$

Note that $Y(t)$ is independent of δ. It follows that

$$\varepsilon(t) - \delta v(t) = g(t),$$

where $g(t)$ is $o(\delta) + O(\varepsilon(t)^2)$ and continuous, and $g'(t)$ is $o(\delta) + O(\varepsilon(t)^2)$ and piecewise continuous, giving the desired result. □

Theorem 2.1 shows that TP as defined by A is equivalent to causing the interpolant to solve a nearby system of differential equations. Asymptotically, the difference between the original system and the nearby system must depend linearly upon δ. Genuine TP also requires that the function $v(t)$ in A should not be everywhere zero. A glance at the proof of Theorem 2.1 shows that this will be the case if $\gamma(t) \not\equiv 0$ in B. We should also note that A is quite a strong condition in the sense that the first derivative of $\eta(t)$ must also have a global error that varies linearly with δ, and the limit function, $v'(t)$, must be continuous. First derivative global errors will be considered in Section 5.

We mentioned in the previous section that, ideally, the global error should vary linearly over the set of all 'reasonable' tolerances. Theorem 2.1 is an asymptotic result ($\delta \to 0$) and hence only comes into effect when δ is sufficiently small (with 'sufficiently small' being a problem-dependent concept). One purpose of the numerical experiments described in the next section is to see the extent to which

useful practical conclusions can be drawn from this type of asymptotic analysis. The theorem also contains the implicit assumption that no rounding errors are made in the computations—it is clear that with finite-precision arithmetic the global error cannot decrease indefinitely with δ.

THEOREM 2.2 *If conditions* A *and* B *defined in Theorem* 2.1 *hold, then the local error at each meshpoint satisfies*

$$\mathrm{le}_n = \gamma(t_n)h_n\delta + o(h_n\delta).$$

Proof. Let $\varepsilon_n(t)$ denote the local error in $\eta(t)$ over $[t_{n-1}, t_n]$; that is, $\varepsilon_n(t) := \eta(t) - z_n(t)$, where $z_n(t)$ is the local solution for the step. Condition B then says that

$$z_n'(t) + \varepsilon_n'(t) - f\big(t, z_n(t) + \varepsilon_n(t)\big) = \gamma(t)\delta + s(t), \qquad \varepsilon_n(t_{n-1}) = 0.$$

Expanding the term $f\big(t, z_n(t) + \varepsilon_n(t)\big)$, and noting that $z_n'(t) = f\big(t, z_n(t)\big)$, we find

$$\varepsilon_n'(t) - f_y\big(t, z_n(t)\big)\varepsilon_n(t) = \gamma(t)\delta + \hat{s}(t), \qquad \varepsilon_n(t_{n-1}) = 0,$$

where $\hat{s}(t)$ is piecewise continuous and $o(\delta) + O(\varepsilon_n(t)^2)$. We may replace $z_n(t)$ with the true solution $y(t)$ in the above expression to give

$$\varepsilon_n'(t) - f_y\big(t, y(t)\big)\varepsilon_n(t) = \gamma(t)\delta + \bar{s}(t), \qquad \varepsilon_n(t_{n-1}) = 0, \tag{2.2}$$

where $\bar{s}(t)$ is $o(\delta) + O\big(\varepsilon_n(t)^2\big)$. Our aim is to find an expression for $\varepsilon_n(t_n) = \mathrm{le}_n$. The solution of (2.2) at t_n has the form (c.f. the proof of Theorem 2.1)

$$\varepsilon_n(t_n) = Y(t_n)\int_{t_{n-1}}^{t_n} Y^{-1}(t)\{\gamma(t)\delta + \bar{s}(t)\}\,\mathrm{d}t,$$

where $Y(t)$ is the fundamental solution matrix. Hence,

$$\varepsilon_n(t_n) = \left\{Y(t_n)\frac{1}{h_n}\int_{t_{n-1}}^{t_n} Y^{-1}(t)\gamma(t)\,\mathrm{d}t\right\}h_n\delta + Y(t_n)\int_{t_{n-1}}^{t_n} Y^{-1}(t)\bar{s}(t)\,\mathrm{d}t. \tag{2.3}$$

A similar expression can be obtained for any $t \in (t_{n-1}, t_n)$, showing that $\varepsilon_n(t)$ is $O(h_n\delta)$ over $[t_{n-1}, t_n]$. It follows that $\bar{s}(t)$ is $o(\delta)$ and hence the second term on the right-hand side of (2.3) is $o(h_n\delta)$. Since $Y^{-1}(t)\gamma(t)$ is continuous, we have

$$\frac{1}{h_n}\int_{t_{n-1}}^{t_n} Y^{-1}(t)\gamma(t)\,\mathrm{d}t = Y^{-1}(t_n)\gamma(t_n) + o(1).$$

Hence (2.3) becomes

$$\varepsilon_n(t_n) = \gamma(t_n)h_n\delta + o(h_n\delta),$$

as required □

We see from Theorem 2.2 that a consequence of TP (in the sense of A) is that each component of the local-error-per-unit-step must depend linearly upon δ. Moreover, at each meshpoint the local-error-per-unit-step must be asymptotically equal to the *same* multiple of δ as the defect.

THEOREM 2.3 *Let $\eta(t)$ be the ideal interpolant. Then conditions* A *and* B *defined in Theorem* 2.1 *hold if and only if the local error is controlled in such a way that*

$$\mathrm{le}_n = \gamma(t_n)h_n\delta + o(h_n\delta). \tag{2.4}$$

Proof. To prove the 'if' part of the theorem, we show that the type of local error control above causes $\eta_I(t)$ to satisfy condition B.

From the definition of $\eta_I(t)$ in (2.1) we have

$$\begin{aligned}\eta_I'(t) - f\big(t, \eta_I(t)\big) &= z_n'(t) + \frac{\mathrm{le}_n}{h_n} - f\left(t, z_n(t) + \frac{(t-t_{n-1})}{h_n}\mathrm{le}_n\right)\\ &= z_n'(t) + \frac{\mathrm{le}_n}{h_n} - f\big(t, z_n(t)\big) + O(\mathrm{le}_n)\\ &= \frac{\mathrm{le}_n}{h_n} + O(\mathrm{le}_n).\end{aligned} \tag{2.5}$$

So, under our assumption about the local error control, we have

$$\eta_I'(t) - f\big(t, \eta_I(t)\big) = \gamma(t_n)\delta + o(\delta).$$

The continuity of $\gamma(t)$ ensures that $\gamma(t_n) = \gamma(t) + o(1)$ for any $t \in [t_{n-1}, t_n]$, and hence

$$\eta_I'(t) - f\big(t, \eta_I(t)\big) = \gamma(t)\delta + o(\delta).$$

Since $\eta_I(t)$ is piecewise C^1, the $o(\delta)$ term above must be piecewise continuous as required for condition B.

The converse of the theorem is simply a special case of Theorem 2.2. □

Theorem 2.3 shows that the relationship between the local-error-per-unit-step and the defect is not only necessary for A and B, but is also sufficient in the case where $\eta(t)$ is the ideal interpolant. It also follows from Theorems 2.2 and 2.3 that if conditions A and B hold for any interpolant then they also hold for the ideal interpolant.

The next result is a corollary to Theorem 2.3 that categorizes a general class of error control schemes for which A and B hold with the ideal interpolant.

COROLLARY 2.1 *Suppose that on every step we ensure that*

$$\|E(y_{n-1}, t_{n-1}, h_n)\| = \delta + o(\delta). \tag{2.6}$$

Here $\|\bullet\|$ denotes a weighted norm, whose weights may depend upon y_{n-1} and y_n (as discussed in Section 1*) and E is a continuous function of the form*

$$E(y_{n-1}, t_{n-1}, h_n) = \hat{\psi}(y_{n-1}, t_{n-1})h_n^p + O(h_n^{p+1}), \tag{2.7}$$

where $\hat{\psi}$ is continuous, independent of h_n, and satisfies $\|\hat{\psi}\big(y(t), t\big)\| \neq 0$ on $[t_0, t_{\mathrm{out}}]$. (In the case where the weights in $\|\bullet\|$ depend upon the numerical solution, we assume that $\|\hat{\psi}\big(y(t), t\big)\| \neq 0$ on $[t_0, t_{\mathrm{out}}]$ for all sufficiently small δ.) Then the ideal interpolant satisfies conditions A and B defined in Theorem 2.1.

Proof. We will show that the type of error control given by (2.6) implies condition (2.4).

From (2.7) we have

$$\frac{\|E(y_{n-1}, t_{n-1}, h_n)\|}{\|\hat{\psi}(y_{n-1}, t_{n-1})\|} = h_n^p + O(h_n^{p+1}). \tag{2.8}$$

Now, as indicated in Section 1, the local error in the pth order Runge–Kutta formula is known to satisfy

$$\mathrm{le}_n = \psi(y_{n-1}, t_{n-1})h_n^{p+1} + O(h_n^{p+2}).$$

Using (2.8) this may be written

$$\mathrm{le}_n = \psi(y_{n-1}, t_{n-1}) \frac{\|E(y_{n-1}, t_{n-1}, h_n)\|}{\|\hat{\psi}(y_{n-1}, t_{n-1})\|} h_n + O(h_n^{p+2}).$$

The local error control (2.6) then gives

$$\mathrm{le}_n = \frac{\psi(y_{n-1}, t_{n-1})}{\|\hat{\psi}(y_{n-1}, t_{n-1})\|} \delta h_n + O(h_n^{p+2}) + o(h_n\delta). \tag{2.9}$$

We now have an expression for the local error that is almost of the form required by Theorem 2.3. A minor technicality is that the weighted norm $\|\bullet\|$ depends upon the numerical solution and hence upon δ. However, if we define $|||\bullet|||$ to be the corresponding weighted norm with true solution values $y(t_{n-1})$, $y(t_n)$, rather than y_{n-1}, y_n, appearing in the weights, then

$$|||\hat{\psi}(y_{n-1}, t_{n-1})||| = \|\hat{\psi}(y_{n-1}, t_{n-1})\| \big(1 + O(\varepsilon(t_{n-1})) + O(\varepsilon(t_n))\big),$$

where we recall that $\varepsilon(t_{n-1}) = y_{n-1} - y(t_{n-1})$ and $\varepsilon(t_n) = y_n - y(t_n)$. Since the stepsizes tend to zero as $\delta \to 0$, it follows from standard theory (see, for example, Hairer, Nørsett, & Wanner, 1987: Thm 3.4, p. 160) that the global error also tends to zero as $\delta \to 0$. So

$$|||\hat{\psi}(y_{n-1}, t_{n-1})||| = \|\hat{\psi}(y_{n-1}, t_{n-1})\| \big(1 + o(1)\big).$$

Hence, from (2.9),

$$\mathrm{le}_n = \frac{\psi(y_{n-1}, t_{n-1})}{|||\hat{\psi}(y_{n-1}, t_{n-1})|||} \delta h_n + O(h_n^{p+2}) + o(h_n\delta). \tag{2.10}$$

The continuity of ψ and $\hat{\psi}$ allows us to replace y_{n-1} and t_{n-1} by $y(t_n)$ and t_n respectively in the right-hand side of (2.10). Now it is clear from (2.6) and (2.7) that an $O(h_n^{p+2})$ term must also be $O(h_n^2\delta)$ and hence must be $o(h_n\delta)$. Thus (2.4) holds with $\gamma(t) = \psi(y(t), t)/|||\hat{\psi}(y(t), t)|||$. □

Roughly, the corollary tells us that we can achieve A by controlling a smoothly varying $O(h_n^p)$ quantity on every step. Considering local error control, the EPUS mode has $\mathrm{est}_n := \|y_n - \bar{y}_n\|/h_n = O(h_n^p)$, and if we are in the usual situation where the two formulae in the RK pair have orders that differ by one, then XEPS gives $\mathrm{est}_n := \|y_n - \bar{y}_n\| = O(h_n^p)$. In both cases it follows from (1.3) that the local error estimate has an expansion of the form (2.7). Also, defect control based on a higher order interpolant has $\mathrm{est}_n = O(h_n^{l-1}) = O(h_n^p)$ and (1.5) gives the required expansion.

The condition (2.6) is satisfied when the standard asymptotically based stepsize selection formula is used. After a step from t_{n-1} to $t_{n-1}+h_n$, the formula is

$$h_{new} = \theta h_{opt}, \qquad h_{opt} = \left(\frac{\delta}{\text{est}_n}\right)^{1/p} h_n. \tag{2.11}$$

Here h_{opt} is the asymptotically optimal stepsize, and the constant safety factor $\theta \in (0,1)$ is introduced to reduce the chance of the step being rejected. The formula can be used to compute a new stepsize after a successful step, or after a rejected step. (In the latter case some authors favour a more crude strategy such as halving the stepsize.) Strictly, (2.11) implies that $\text{est}_n = \theta^p \delta + o(\delta)$, so we must interpret $\|E(y_{n-1}, t_{n-1}, h_n)\|$ in (2.6) as $\theta^{-p}\,\text{est}_n$. We then have the following corollary.

COROLLARY 2.2 *Suppose that one of the following error control modes is used*

(i) EPUS
(ii) XEPS *with a* $(p-1)$th *and* pth *order pair of formulae*
(iii) *defect control with a higher-order interpolant*

with stepsize selection based on the formula (2.11). *Then, provided that* $\hat{\psi}$ *in* (2.7) *satisfies* $\|\hat{\psi}(y(t),t)\| \neq 0$ *on* $[t_0, t_{\text{out}}]$ *for all sufficiently small* δ, *and the initial stepsize is chosen so that* (2.6) *holds on the first step, the ideal interpolant satisfies conditions* A *and* B *defined in Theorem* 2.1.

Each of the three error control modes listed in Corollary 2.2 will be implemented in the numerical tests of the next section.

As a final point, we mention that some nonasymptotic theory that is relevant for stiff and mildly stiff initial value problems was developed by Hall (1985, 1986) and by Hall & Higham (1988) and Higham & Hall (1989). This analysis gives algebraic conditions involving the RK coefficients and the dominant eigenvalue(s) of the local Jacobian for determining whether or not a smooth stepsize sequence will arise. The same conditions also determine whether or not TP will be observed.

3. Discrete formulae

In this section we describe numerical tests on a variety of discrete RK formulae and error control techniques. The following formula pairs were tested: a 3,4 pair of Nørsett that appears in Enright *et al.* (1986), the 4,5 pair RK5(4)7FM of Dormand & Prince (1980), and a 7,8 pair of Sharp & Smart (1989). Each pair was implemented in EPUS and XEPS mode. We also tested two defect control techniques based on the 4,5 pair. The first one uses the locally $O(h_n^6)$ Dormand–Prince–Shampine (DPS) interpolant of Shampine (1986). This interpolant is constructed in such a way that the defect is always zero at the midpoint of a step, hence $\tau^* = \frac{1}{2}$ should not be used as a sample point. It is also reasonable to ask for the sampled value to reflect the maximum size of the defect over the step. Although, in general, the shape of the defect will depend on the differential

equations and will vary from step to step, in Higham (1990: Appendix B) we use an asymptotic expansion of the defect to justify the choice $\tau^* = 0{\cdot}87832$. The second defect control scheme that we tested is based on the more expensive locally $O(h_n^6)$ Hermite–Birkhoff (HB) interpolant from Higham (1989*b*). In this case the sample point $\tau^* = 0{\cdot}89994$ is guaranteed to control the maximum defect, asymptotically.

For stepsize selection, we used the formula (2.11) with a safety factor of 0·9, subject to the restriction

$$\tfrac{1}{10}h_{\text{old}} \leq h_{\text{new}} \leq 5h_{\text{old}}.$$

The final stepsize was further restricted so that the endpoint t_{end} was reached exactly.

Our testing made use of the nonstiff component of the DETEST package (Enright & Pryce, 1987). Detailed results for the 25 smooth scaled problems are given in Higham (1990). Here, in Tables 3.1 and 3.2, we present the ratio of the endpoint global error and the tolerance for two typical problems. Uniform absolute weights were used, with tolerances of $10^{-2}, 10^{-3}, \ldots, 10^{-10}$. Ideally, for each method and test problem this ratio will remain constant over all tolerances. Here, and throughout the numerical testing, the infinity norm was used.

A direct measure of the TP of a method on each problem is also available from DETEST. Based on the model

$$\|\text{global error}\| \approx C \times \delta^E, \tag{3.1}$$

a least-squares fit is found by minimizing

$$R = \sum_{i=1}^{\text{NTOL}} [A + E \times \log_e(\delta_i) - \log_e \|\text{global error}_i\|]^2,$$

TABLE 3.1
Endpoint global error divided by tolerance for Problem C5 of DETEST

δ	EPUS34	EPUS45	EPUS78	XEPS34	XEPS45	XEPS78	DPSdef	HBdef
10^{-2}	10·21	72·22	0·79	0·87	13·26	0·08	5·76	4·63
10^{-3}	10·73	109·29	2·57	0·86	31·36	0·07	9·08	6·85
10^{-4}	10·73	207·71	2·95	0·83	37·93	0·05	8·61	6·82
10^{-5}	10·43	102·02	4·42	0·81	22·85	0·06	8·48	6·85
10^{-6}	10·21	31·15	9·07	0·80	13·79	0·04	8·94	6·46
10^{-7}	10·11	11·28	10·71	0·79	8·60	0·04	7·84	5·42
10^{-8}	10·06	5·43	12·31	0·78	6·54	0·05	8·11	5·18
10^{-9}	10·04	4·88	13·53	0·77	5·17	0·06	8·07	4·81
10^{-10}	10·02	4·67	14·48	0·77	4·29	0·07	7·75	4·44
E (edpt)	1·003	1·215	0·853	1·007	1·104	1·008	0·995	1·015
RES (edpt)	7E−3	2E−1	1E−1	4E−3	1E−1	1E−1	5E−2	6E−2
E (max)	1·001	1·218	0·853	1·007	1·103	0·979	0·995	1·014
RES (max)	7E−3	2E−1	1E−1	5E−3	1E−1	1E−1	5E−2	6E−2

TABLE 3.2
Endpoint global error divided by tolerance for Problem E4 of DETEST

δ	EPUS34	EPUS45	EPUS78	XEPS34	XEPS45	XEPS78	DPSdef	HBdef
10^{-2}	3·15	0·20	3·40	0·11	0·01	0·06	0·01	0·01
10^{-3}	2·01	0·43	4·27	0·09	0·04	0·01	0·11	0·83
10^{-4}	2·73	2·27	0·65	0·11	0·03	0·00	1·23	0·15
10^{-5}	2·72	1·78	0·82	0·12	0·07	0·01	0·31	1·64
10^{-6}	2·95	2·52	1·90	0·13	0·10	0·01	0·82	0·78
10^{-7}	3·24	2·35	1·61	0·14	0·14	0·02	0·41	0·80
10^{-8}	3·26	2·52	1·19	0·14	0·16	0·04	0·50	0·73
10^{-9}	3·31	2·71	1·32	0·14	0·18	0·08	0·52	0·81
10^{-10}	3·32	2·88	1·55	0·14	0·19	0·09	0·52	0·76
E (edpt)	0·984	0·879	1·035	0·978	0·855	0·914	0·870	0·857
RES (edpt)	5E−2	2E−1	2E−1	3E−2	1E−1	3E−1	4E−1	5E−1
E (max)	0·997	0·920	1·035	0·993	0·910	0·994	0·925	0·922
RES (max)	3E−2	2E−1	1E−1	4E−2	2E−1	2E−1	6E−1	3E−1

where δ_i, $i = 1,\ldots,$ NTOL, is the range of tolerances used. The values of the proportionality constant $C := \exp(A)$, the exponent E, and the root mean square residual RES $:= (R/\text{NTOL})^{\frac{1}{2}}$ are returned. A method which comes close to exhibiting TP will have an exponent $E \approx 1$ and a small residual RES. Tables 3.1 and 3.2 also include the exponent and residual values for both the endpoint global error (edpt) and the maximum global error over all meshpoints (max). Although it is clear that asking for linear proportionality of the norm of the endpoint or maximum global errors is a weaker requirement than A in Theorem 2.1, the DETEST results will certainly give insights into the likelihood of a condition such as A being satisfied.

Based on Tables 3.1 and 3.2, and the more detailed results in Higham (1990), we make the following observations on the ratio of the endpoint global error and the tolerance:

(i) On all problems the ratio improves (varies less) as the tolerance decreases. Typically the behaviour is poor for $\delta \geq 10^{-4}$, although the δ for which the ratio begins to settle down depends on the method and the problem.
(ii) For both the EPUS and XEPS modes the relative performance worsens as the order increases—the lower-order methods exhibit better proportionality over a wider range of tolerances.
(iii) For each method and almost all problems, the ratio varies by less than a factor of 1·5 at the most stringent tolerances.
(iv) The performance of the defect control methods DPSdef and HBdef is similar to that of the XEPS45 method. (Note that these methods advance with the same RK formula while controlling different $O(h_n^5)$ quantities.)

The first observation is not surprising, given the asymptotic nature of the analysis in the previous section. With regard to the second observation, Stetter (1980) notes that if an $O(h_n^p)$ quantity is controlled by δ, then the stepsize h_n will

decrease like $O(\delta^{1/p})$. So, in the proofs that lead up to Corollary 2.2, discarding terms with an extra factor of h_n is likely to be much less realistic than discarding terms with an extra factor of δ, and the theory becomes less meaningful as the order p increases. In fact many of the DETEST problems are quite 'easy' in the sense that they are integrated with very few steps, even at quite stringent tolerances. For example, on problem E5 for $\delta \geq 10^{-6}$, all methods required less than 35 steps. (The range of integration is $[0, 20]$ for each DETEST problem.) As we would expect, the stepsizes used for a given problem and tolerance tend to increase with the order of the error estimate est_n. On B2 at $\delta = 10^{-8}$, for example, the XEPS34 method uses 267 steps while XEPS78 uses a mere 23.

An interesting side issue that is worth mentioning is the variation of the global error across different methods for a fixed problem and tolerance. Some insight can be gained by examining the proofs of Theorems 2.1 and 2.2. If condition A of Theorem 2.1 holds, then $v(t)$ solves

$$v'(t) - f_y(t, y(t))v(t) = \gamma(t), \qquad v(t_0) = 0. \tag{3.2}$$

The right-hand side $\gamma(t)$ is closely related to the local-error-per-unit-step, le_n/h_n. Theorem 2.2 tells us that, at each meshpoint, $\gamma(t)\delta$ and le_n/h_n are asymptotically equal. However, it is important to note that $\gamma(t) : \mathbb{R} \to \mathbb{R}^N$ is a vector-valued function of t in general. Hence, even if two different methods both keep the *norm* of le_n/h_n equal to δ on each step, the individual components may have different signs and magnitudes, and hence the solutions $v(t)$ of (3.2) will not necessarily be close. On the other hand, it can be argued that, of the different error control types, EPUS should be the most 'method-insensitive' since with other strategies a t-dependent multiple of the local-error-per-unit-step is controlled, and hence $\gamma(t)$ has a much greater degree of freedom. This is borne out in the results; the variation across the columns is generally less dramatic with EPUS than with XEPS.

4. Interpolants

We now turn our attention to the use of interpolants to provide continuous approximations to $y(t)$. The first question that we ask is whether any of the standard interpolants can satisfy condition B (and hence condition A) of Theorem 2.1. The answer is no for any of the C^1 schemes that satisfy $q'(t_n) = f(t_n, y_n)$. For such schemes the defect is forced to be zero at the meshpoints, and since the location of the meshpoints varies with δ, condition B cannot hold. More specifically, for $t = t_{n-1} + \tau h_n \in (t_{n-1}, t_n]$ any continuous extension of the form (1.4) has a local error of the form

$$q(t_{n-1} + \tau h_n) - z_n(t_{n-1} + \tau h_n) = h_n^l \sum_{j=1}^{m} a_j(\tau) F_j(t_{n-1}, y_{n-1}) + O(h^{l+1}), \tag{4.1}$$

and a defect of the form

$$q'(t_{n-1} + \tau h_n) - f\big(t_{n-1} + \tau h_n, q(t_{n-1} + \tau h_n)\big) = h_n^{l-1} \sum_{j=1}^{m} a_j'(\tau) F_j(t_{n-1}, y_{n-1}) + O(h^l). \tag{4.2}$$

Here $F_j(t_{n-1}, y_{n-1})$ is an elementary differential of f, and $a_j(\tau)$ is a scalar polynomial that is independent of f. Because of the presence of the $a_j'(\tau)$ factors, the value of the defect at the point t will depend on the relative location of t and the meshpoints t_{n-1}, t_n. Again, since the mesh varies with δ, condition B cannot be satisfied. Although this is discouraging, it is not the whole story, since Theorem 2.1 asks for proportionality of the global error in the interpolant *and* its first derivative. As we show below, useful results about the global error can be obtained by comparing $q(t)$ with $\eta_1(t)$.

We suppose now that an error control of the type defined in Corollary 2.1 is used, so that the ideal interpolant satisfies

$$\eta_1(t) - y(t) = v(t)\delta + g(t), \tag{4.3}$$

with $v(t)$ and $g(t)$ as in Theorem 2.1. For $t \in (t_{n-1}, t_n]$, it is helpful to split the global error in $q(t)$ into three parts:

$$q(t) - y(t) = [q(t) - z_n(t)] - [\eta_1(t) - z_n(t)] + [\eta_1(t) - y(t)]. \tag{4.4}$$

We have $q(t) - z_n(t) = O(h_n^l)$ for the local error in $q(t)$, where $l = p + 1$ or $l = p$ depending upon whether $q(t)$ is chosen to be higher or lower order (see Section 1). The local error in $\eta_1(t)$ satisfies

$$\eta_1(t) - z_n(t) = O(\mathrm{le}_n) = O(h_n^{p+1}) = O(h_n\delta) = o(\delta).$$

Hence (4.4) may be written

$$q(t) - y(t) = v(t)\delta + o(\delta) + O(h_n^l). \tag{4.5}$$

If $q(t)$ is higher order then the $O(h_n^l)$ term is $o(\delta)$, so that (4.5) becomes

$$q(t) - y(t) = v(t)\delta + o(\delta), \tag{4.6}$$

showing that $q(t)$ inherits the same proportionality as $\eta_1(t)$. (Note that (4.6) is weaker than condition A, since A asks for $g'(t)$ to be $o(\delta)$.) A lower-order interpolant contributes an $O(h_n^p) = O(\delta)$ term in (4.5), and hence $v(t)\delta$ will not necessarily dominate. In this case the relative sizes of the $[q(t) - z_n(t)]$ local error term and the $[\eta_1(t) - y(t)]$ global error term in (4.4) will depend upon the differential equations. The expansion (4.1) shows that the local error term has a highly mesh-dependent nature. Hence in general we cannot deduce a result of the form (4.6) for a lower-order interpolant.

To investigate the behaviour of interpolation schemes in practice, we performed numerical tests on four interpolants that have been derived for the Dormand–Prince RK5(4)7FM pair. We used lower-order (locally $O(h_n^5)$) interpolants from Dormand & Prince (1986) and Shampine (1986), and higher-order (locally $O(h_n^6)$) interpolants from Higham (1989*b*) and Shampine (1986), denoted DP[$O(h^5)$], DPS[$O(h^5)$], HB[$O(h^6)$], and DPS[$O(h^6)$] respectively. The 4,5 pair was implemented in XEPS mode and, by partitioning each basic RK step into 10 equally spaced substeps, we used DETEST to compute

$$\max_n \left\{ \max_{i \le i \le 10} \{ \| q(t_{n-1} + \tfrac{1}{10} i h_n) - y(t_{n-1} + \tfrac{1}{10} i h_n) \|_\infty \} \right\} \Big/ \delta$$

TABLE 4.1
Maximum global error divided by tolerance for Problem C5 of DETEST

δ	DP[$O(h^5)$]	DPS[$O(h^5)$]	DPS[$O(h^6)$]	HB[$O(h^6)$]
10^{-2}	17·96	17·96	17·96	17·96
10^{-3}	31·12	31·12	31·12	31·12
10^{-4}	24·46	24·46	24·46	24·46
10^{-5}	21·27	21·27	21·27	21·17
10^{-6}	13·11	13·11	13·11	13·11
10^{-7}	8·83	8·83	8·12	8·12
10^{-8}	6·63	6·59	6·37	6·38
10^{-9}	5·41	5·37	5·14	5·15
10^{-10}	4·59	4·55	4·32	4·32
E	1·103	1·103	1·107	1·107
RES	1·1 E − 1	1·1 E − 1	1·1 E − 1	1·1 E − 1

for each problem. If (4.6) holds, then this ratio should remain constant as δ decreases. Sample results for problems C5 and E4 are given in Tables 4.1 and 4.2. Over the 25 test problems (see Higham, 1990, for full details) we found that

(i) the behaviour of both $O(h_n^6)$ interpolants is reasonable (in fact the proportionality of the maximum global error in the interpolant is as good as that of the discrete formula for each problem);
(ii) the proportionality of the $O(h_n^5)$ interpolants is similar to that of the higher order interpolants on many problems, but is noticeably worse on A5, B3, E1, E4, and E5, and significantly worse on A4; this problem-dependent behaviour is to be expected from the analysis above.

For further illustration, we performed detailed numerical tests on the following

TABLE 4.2
Maximum global error divided by tolerance for Problem E4 of DETEST

δ	DP[$O(h^5)$]	DPS[$O(h^5)$]	DPS[$O(h^6)$]	HB[$O(h^6)$]
10^{-2}	0·30	0·30	0·17	0·07
10^{-3}	0·20	0·20	0·13	0·07
10^{-4}	0·52	0·57	0·18	0·09
10^{-5}	0·77	0·83	0·19	0·10
10^{-6}	1·55	1·63	0·14	0·13
10^{-7}	3·27	3·36	0·16	0·16
10^{-8}	3·69	3·76	0·17	0·17
10^{-9}	4·16	4·23	0·19	0·19
10^{-10}	4·95	4·95	0·19	0·19
E	0·814	0·815	0·989	0·938
RES	1·4 E − 1	1·4 E − 1	5·2 E − 2	3·8 E − 2

problems

(1) A problem of Fehlberg (see Hairer, Nørsett, & Wanner, 1987: p. 174):

$$y_1' = 2ty_1 \log [\max (y_2, 10^{-3})], \qquad y_1(0) = 1,$$
$$y_2' = 2ty_2 \log [\max (y_1, 10^{-3})], \qquad y_2(0) = e, \quad 0 \leq t \leq 5,$$

which has solution $y_1 = \exp (\sin t^2)$, $y_2 = \exp (\cos t^2)$.

(2) Problem A4 (unscaled) from Enright & Pryce (1987):

$$y' = \tfrac{1}{4}y(1 - \tfrac{1}{20}y), \qquad y(0) = 1, \quad 0 \leq t \leq 20,$$

for which

$$y = \frac{20}{1 + 19 \exp (-\frac{1}{4}t)}.$$

We present results for a uniform absolute weighting scheme; experiments with other weighting schemes gave similar qualitative results. As a point of reference, we plot the norm of the meshpoint global error, scaled by δ, for tolerances of 10^{-4}, 10^{-6}, 10^{-8}, and 10^{-10} in Figs 1 and 4. In both cases the global error ratio converges to a discernible limit function.

The corresponding global errors in the higher- and lower-order interpolants of Shampine (1986) are plotted in Figs 2, 3, 5, and 6. These were generated by evaluating the global error at 100 equally spaced points in the interval of integration. For the higher-order interpolant, we see that on the Fehlberg problem the global error ratios have almost exactly the same shape as for the discrete solution. On problem A4, the curves are somewhat 'wobbly', but the behaviour improves as the tolerance decreases. The proportionality of the lower-order interpolant is comparable with that of the discrete formula on the Fehlberg problem, but is slightly erratic over the first part of the integration. On problem A4, however, the lower-order interpolant behaves very poorly; although the global error ratios remain bounded, they are not smooth and do not appear to approach a limit as δ decreases.

The behaviour of the lower-order interpolant on these two problems can be explained by looking at the actual size of the global errors. On the Fehlberg problem, the global error in the lower-order interpolant is the same size as that of the meshpoint solution, which shows that of the two $O(\delta)$ terms $[\eta_1(t) - y(t)]$ and $[q(t) - z_n(t)]$ in (4.4), the first term is dominating in the plots. On problem A4 the meshpoint global error is never more than 1.5δ, and the $[q(t) - z_n(t)]$ term clearly takes over in Fig. 6. We also mention that fairly large stepsizes are used on problem A4—only 9, 19, 45, and 108 steps are needed at $\delta = 10^{-4}$, 10^{-6}, 10^{-8}, and 10^{-10} respectively. Thus, with the higher-order interpolant, we would expect the $O(h_n\delta)$ term $[q(t) - z_n(t)]$ in (4.4) to be significant, even at quite stringent tolerances—this is what appears to be happening in Fig. 5.

Although carried out for a different purpose, numerical testing of several lower- and higher-order interpolants was performed in Enright *et al.* (1986). There the authors used DETEST to compare the maximum global error in an interpolant with the maximum global error in the underlying discrete RK method.

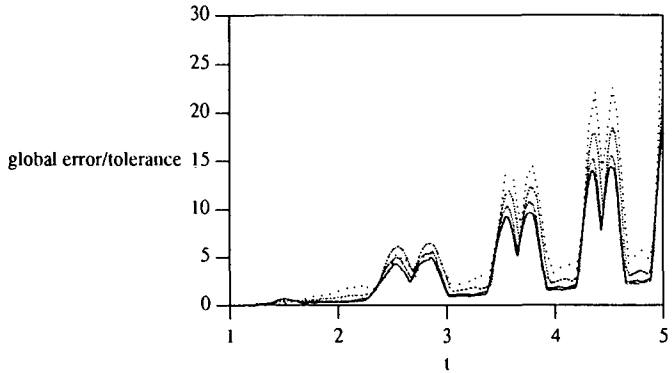

FIG. 1. Fehlberg problem: tolerances 1 E − 4, 1 E − 6, 1 E − 8, 1 E − 10 (lines become more solid as tolerance decreases).

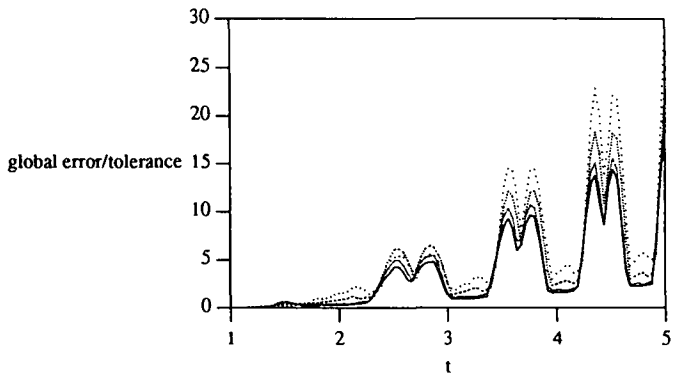

FIG. 2. Fehlberg problem, higher-order interpolant: tolerances 1 E − 4, 1 E − 6, 1 E − 8, 1 E − 10 (lines become more solid as tolerance decreases).

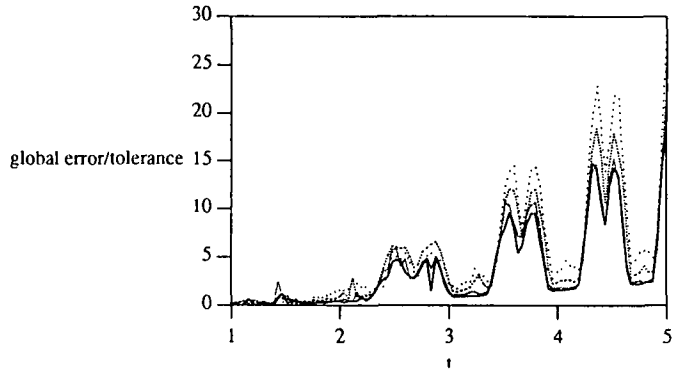

FIG. 3. Fehlberg problem, lower-order interpolant: tolerances 1 E − 4, 1 E − 6, 1 E − 8, 1 E − 10 (lines become more solid as tolerance decreases).

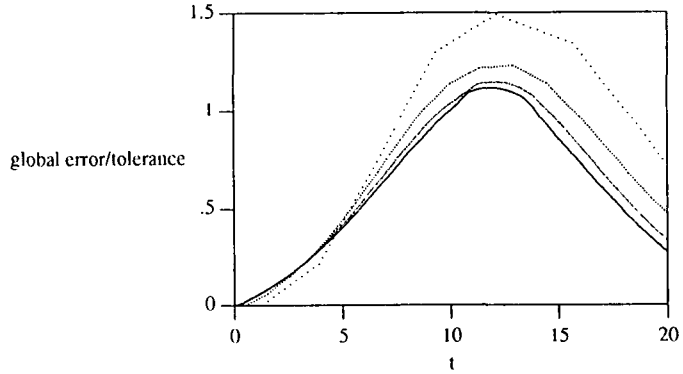

FIG. 4. Problem A4: tolerances 1 E − 4, 1 E − 6, 1 E − 8, 1 E − 10 (lines become more solid as tolerance decreases).

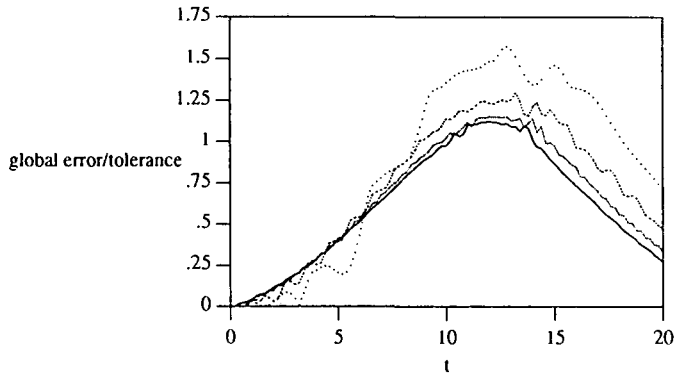

FIG. 5. Problem A4, higher-order interpolant: tolerances 1 E − 4, 1 E − 6, 1 E − 8, 1 E − 10 (lines become more solid as tolerance decreases).

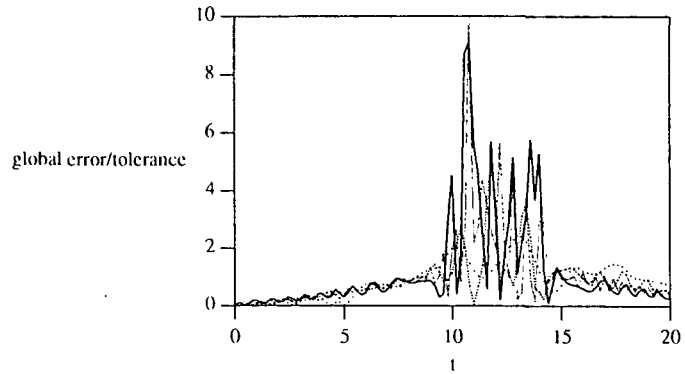

FIG. 6. Problem A4, lower-order interpolant: tolerances 1 E − 4, 1 E − 6, 1 E − 8, 1 E − 10 (lines become more solid as tolerance decreases).

Their results are in broad agreement with ours, and the arguments above help to explain the somewhat erratic behaviour of the lower-order interpolants reported in that paper.

5. First derivatives

In addition to making an interpolant $q(t)$ available, it is common for modern initial value codes to provide access to $q'(t)$ (see, for example, Cash, 1989). In this section we look at the behaviour of the corresponding global error, $q'(t) - y'(t)$.

From Corollary 2.1, we see that if a suitable error control strategy is used, the ideal interpolant $\eta_I(t)$ satisfies condition A of Theorem 2.1 and hence

$$\eta_I'(t) - y'(t) = v'(t)\delta + o(\delta),$$

where $v'(t)$ is independent of δ and continuous. (We point out that although $\eta_I'(t)$ is only piecewise continuous, the jumps at the meshpoints are $O(\mathrm{le}_n) + o(\delta)$ and hence do not affect the leading term in the global error.)

We saw in the previous section that standard computable interpolants $q(t)$ will not satisfy condition A, and hence we cannot deduce anything about the TP of $q'(t)$ directly. However, useful information is obtained indirectly by differentiating (4.4) to give

$$q'(t) - y'(t) = [q'(t) - z_n'(t)] - [\eta_I'(t) - z_n'(t)] + [\eta_I'(t) - y'(t)]. \tag{5.1}$$

It can be shown that for continuous extensions of the form (1.4) the local error contribution $q'(t) - z_n'(t)$ has the same asymptotic leading term as the defect (4.2). So, for fixed t, this term varies with the meshpoint distribution, and hence with δ. For the higher-order interpolants we have $l = p + 1$, so the local error term $q'(t) - z_n'(t)$ is of the same order as the global error term $\eta_I'(t) - y'(t)$. Hence, in general, we cannot expect tolerance proportionality in (5.1). The situation is worse for lower-order interpolants; here $l = p$ and the local error term dominates (5.1) asymptotically.

We mention that instead of using $q'(t)$ to approximate $y'(t)$, the quantity $f(t, q(t))$ could be evaluated. The linearization

$$f(t, q(t)) - f(t, y(t)) = f_y(t, y(t))(q(t) - y(t)) + O([q(t) - y(t)]^2) \tag{5.2}$$

shows that $f(t, q(t))$ inherits TP from $q(t)$. Depending on the application, this approach has two possible weaknesses. First, if a lot of off-meshpoint derivative approximations are needed then the extra f evaluations may prove expensive. Second, at a general point the $y'(t)$ approximation is not the derivative of the $y(t)$ approximation.

To illustrate the behaviour of different first derivative approximations, we give the results of detailed numerical tests on the two problems used in the last section. Again, the XEPS45 pair was used to produce the discrete solution. Figures 7 and 11 plot the global error ratios for meshpoint $f(t_n, y_n)$ values. Results for the ideal interpolant, using 100 equally spaced points in the integration range, are given in Figs 8 and 12. Since the ideal interpolant is not

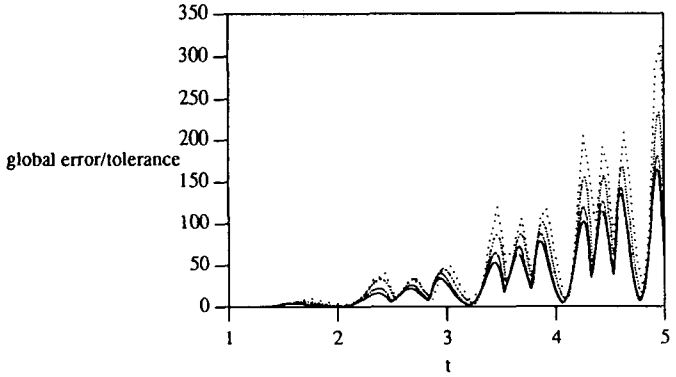

FIG. 7. Fehlberg problem, error in first derivative approximation at meshpoints: tolerances 1 E − 4, 1 E − 6, 1 E − 8, 1 E − 10 (lines become more solid as tolerance decreases).

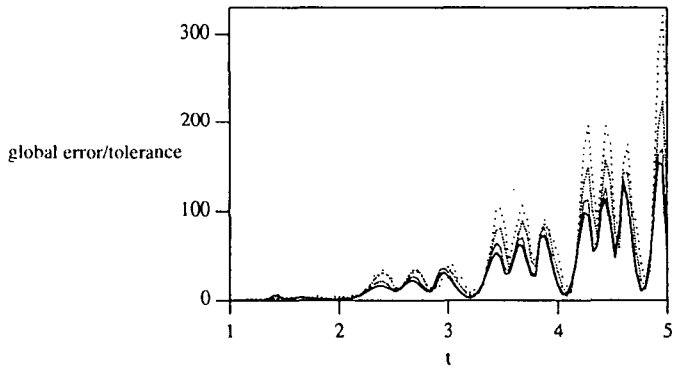

FIG. 8. Fehlberg problem, error in first derivative of ideal interpolant: tolerances 1 E − 4, 1 E − 6, 1 E − 8, 1 E − 10 (lines become more solid as tolerance decreases).

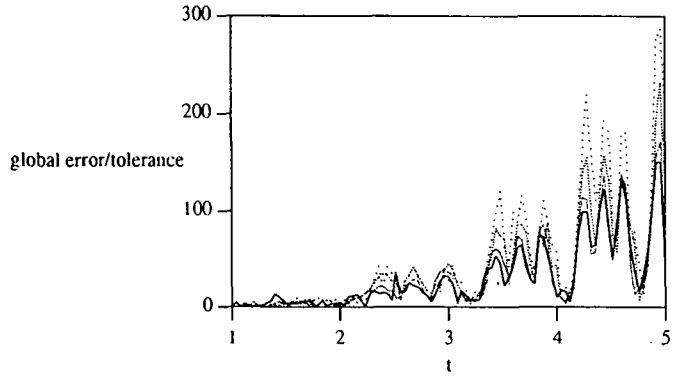

FIG. 9. Fehlberg problem, error in first derivative of higher-order interpolant: tolerances 1 E − 4, 1 E − 6, 1 E − 8, 1 E − 10 (lines become more solid as tolerance decreases).

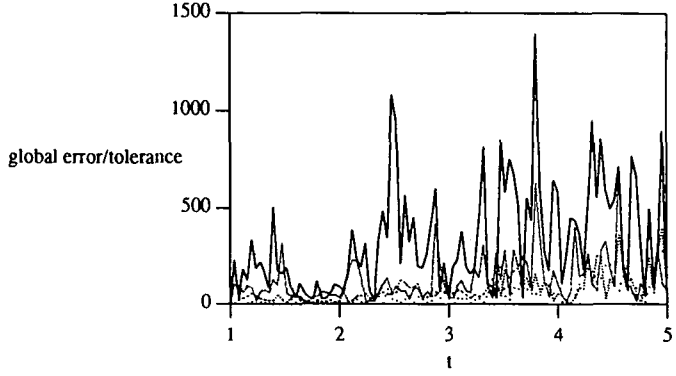

FIG. 10. Fehlberg problem, error in first derivative of lower-order interpolant: tolerances 1 E − 4, 1 E − 6, 1 E − 8, 1 E − 10 (lines become more solid as tolerance decreases).

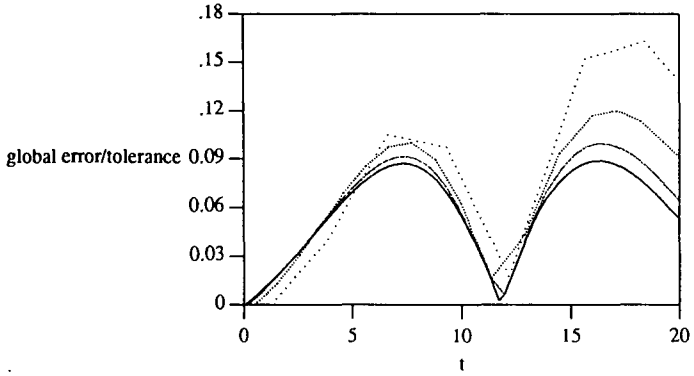

FIG. 11. Problem A4, error in first derivative approximation at meshpoints: tolerances 1 E − 4, 1 E − 6, 1 E − 8, 1 E − 10 (lines become more solid as tolerance decreases).

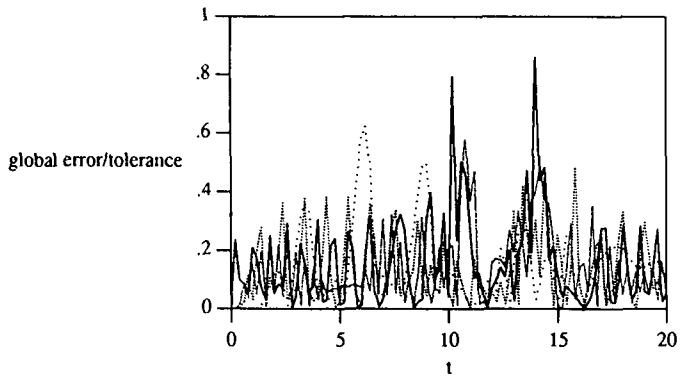

FIG. 13. Problem A4, error in first derivative of higher-order interpolant: tolerances 1 E − 4, 1 E − 6, 1 E − 8, 1 E − 10 (lines become more solid as tolerance decreases).

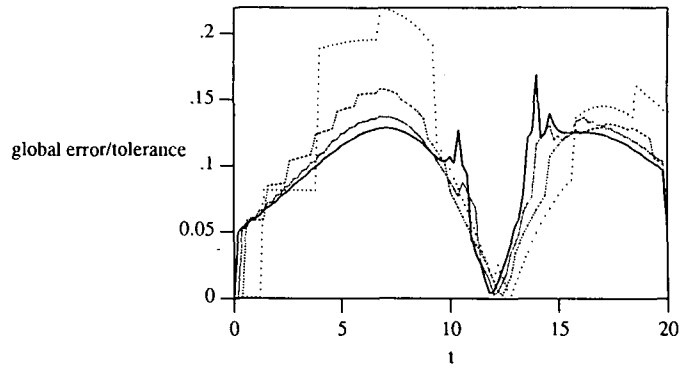

FIG. 12. Problem A4, error in first derivative of ideal interpolant: tolerances 1 E − 4, 1 E − 6, 1 E − 8, 1 E − 10 (lines become more solid as tolerance decreases).

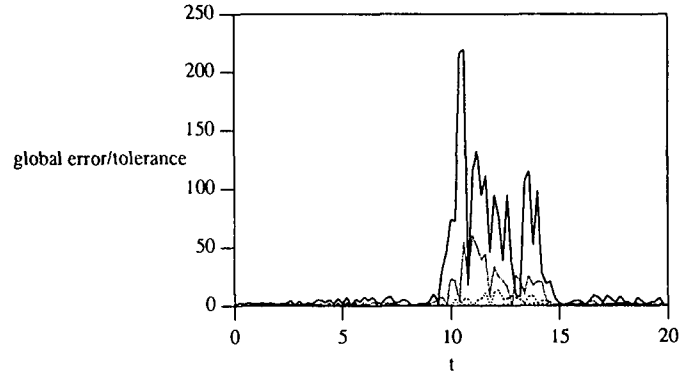

FIG. 14. Problem A4, error in first derivative of lower-order interpolant: tolerances 1 E − 4, 1 E − 6, 1 E − 8, 1 E − 10 (lines become more solid as tolerance decreases).

computable, we used an eighth order formula to obtain the 'true' local solution values. More precisely, we computed the following approximation to $\eta_1'(t) := z_n'(t) + \mathrm{le}_n/h_n$ for $t \in (t_{n-1}, t_n]$:

$$f(t, \bar{u}(t)) + \frac{\widetilde{\mathrm{le}}_n}{h_n}.$$

Here $\bar{u}(t)$ is the result of a step from $\{t_{n-1}, y_{n-1}\}$ of length $t - t_{n-1}$ using an eighth-order RK formula, and $\widetilde{\mathrm{le}}_n = y_n - \bar{u}(t_n)$. The corresponding results for the higher- and lower-order DPS interpolants are plotted in Figs 9 and 13 and Figs 10 and 14 respectively.

We see from the figures that the global error to tolerance proportionality for both the meshpoint derivative and the ideal interpolant derivative is comparable with that of the discrete formula (see Figs 1 and 4). (The $\eta_1'(t)$ behaviour is somewhat erratic on problem A4, but improves as the tolerance decreases.) The higher-order interpolant performs quite well on the Fehlberg problem, except in regions where the global error is small. On problem A4 the global error ratios for the higher-order interpolant remain bounded, but do not settle down to a limit. This behaviour, which we observed in the previous section for the zeroth derivative of the lower-order interpolant, agrees with the theoretical analysis—the two terms $q'(t) - z_n'(t)$ and $\eta_1'(t) - y'(t)$ have the same asymptotic order, and if the first term is relatively large then we will not observe TP. The lower-order interpolant, with its asymptotically dominant $q'(t) - z_n'(t)$ term, fares much worse. On problem A4 the global error to tolerance ratio grows significantly as δ decreases.

6. Conclusions

Tolerance proportionality clearly deserves to vie for a place among the many conflicting objectives of a one-pass integrator. In this work we focused on the potential for TP within existing explicit RK algorithms. The theoretical results presented in Section 2, which are based on some earlier work of Stetter, show that, ignoring rounding errors and with sufficiently small tolerances, a discrete RK formula will exhibit TP if either an EPUS or XEPS local error control method, or a defect control method based on a higher-order interpolant, is suitably implemented. Our numerical tests using the norm of the endpoint global error on the DETEST problems were in agreement with this result. However, we observed that with less stringent tolerances the global error behaviour can be much more erratic—here 'large stepsizes' render the asymptotic results inapplicable. In general, higher-order formulas fare worst because they tend to use larger stepsizes for a given problem and tolerance. Interestingly, the tests of Sharp (1988) found a high-quality 7,8 pair to be more cost-effective than other, lower order pairs. It seems that choosing an RK formula pair involves a trade-off between efficiency (cost versus accuracy achieved) and reliability (in the sense of TP). Our results revealed little difference in the TP performance of EPUS, XEPS, and defect control schemes of a similar order.

New results for continuously extended RK formulae were also presented. We showed that the approximations generated by 'higher order' interpolants (whose local errors have the same asymptotic order as the RK formula) automatically inherit the asymptotic TP properties of the discrete method. The same cannot be said for 'lower order' interpolants; here the local errors have an unpredictable problem-dependent effect on the global error proportionality.

The situation is worse when first derivative approximations are required. In this case higher-order interpolants cannot be guaranteed to behave smoothly, and lower order interpolants will never achieve TP, asymptotically. A simple, but potentially expensive fix is to evaluate f along the interpolant.

Acknowledgements

This paper benefited from the comments of Nick Higham, Ken Jackson and Philip Sharp. The author was supported in this work by the Information Technology Research Centre of Ontario and the Natural Science and Engineering Research Council of Canada.

References

ASCHER, U. M., MATTHEIJ, R. M. M., & RUSSELL, R. D. 1988 *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations.* New Jersey: Prentice-Hall.

CASH, J. R. 1989 A block 6(4) Runge–Kutta formula for nonstiff initial value problems. *ACM Trans. Math. Soft.* **15,** 15–28.

DORMAND, J. R., & PRINCE, P. J. 1980 A family of embedded Runge–Kutta formulae. *J. of Computational and Applied Maths.* **6,** 19–26.

DORMAND, J. R., & PRINCE, P. J. 1986 Runge–Kutta triples. *Comp. and Maths. with Appls.* **12,** 1007–1017.

ENRIGHT, W. H. 1989a A new error-control for initial value solvers. *Applied Maths. and Computation* **31,** 288–301.

ENRIGHT, W. H. 1989b Analysis of error control strategies for continuous Runge–Kutta methods. *SIAM J. Numer. Anal.* **26,** 588–599.

ENRIGHT, W. H., & PRYCE, J. D. 1987 Two FORTRAN packages for assessing initial value methods. *ACM Trans. Math. Soft.* **13,** 1–27.

ENRIGHT, W. H., JACKSON, K. R., NØRSETT, S. P., & THOMSEN, P. G. 1986 Interpolants for Runge–Kutta formulas. *ACM Trans. Math. Soft.* **12,** 193–218.

HAIRER, E., NØRSETT, S. P., & WANNER, G. 1987 *Solving Ordinary Differential Equations I.* Berlin: Springer-Verlag.

HALL, G. 1985 Equilibrium states of Runge–Kutta schemes. *ACM Trans. Math. Soft.* **11,** 289–301.

HALL, G. 1986 Equilibrium states of Runge–Kutta schemes: part II. *ACM Trans. Math. Soft.* **12,** 183–192.

HALL, G., & HIGHAM, D. J. 1988 Analysis of stepsize selection schemes for Runge–Kutta codes. *IMA J. Numer. Anal.* **8,** 305–310.

HIGHAM, D. J. 1989a Robust defect control with Runge–Kutta schemes. *SIAM J. Numer. Anal.* **26,** 1175–1183.

HIGHAM, D. J. 1989b Runge–Kutta defect control using Hermite–Birkhoff interpolation. Technical Report No. 221/89, Department of Computer Science, University of Toronto; to appear in *SIAM J. Sci. Stat. Comput.*

HIGHAM, D. J. 1990 Global error versus tolerance for explicit Runge–Kutta methods. Technical Report No. 233/90, Department of Computer Science, University of Toronto.

HIGHAM, D. J., & HALL, G. 1989 Runge–Kutta equilibrium theory for a mixed relative/absoute error measure. Technical Report No. 218/89, Department of Computer Science, University of Toronto; to appear in the *Proceedings of the 1989 IMA Conference on Computational* ODEs (J. R. Cash and I. Gladwell, Eds).

SHAMPINE, L. F. 1977 Local error control in codes for ordinary differential equations. *Applied Maths. and Computation* **3,** 189–210.

SHAMPINE, L. F. 1985 Interpolation for Runge–Kutta methods. *SIAM J. Numer. Anal.* **22,** 1014–1027.

SHAMPINE, L. F. 1986 Some practical Runge–Kutta formulas. *Math. Comp.* **46,** 135–150.

SHAMPINE, L. F., & WATTS, H. A. 1979 DEPAC—Design of a user oriented package of ODE solvers. Report SAND79-2374, Sandia National Laboratories, Albuquerque, New Mexico.

SHARP, P. W. 1988 Numerical comparisons of explicit Runge–Kutta pairs of orders four through eight. Technical Report 216/88, Department of Computer Science, University of Toronto; to appear in *ACM Trans. Math. Soft.*

SHARP, P. W., & SMART, E. 1989 Private communication.

STETTER, H. J. 1978 Considerations concerning a theory for ODE-solvers. In: *Numerical Treatment of Differential Equations:* Proc. Oberwolfach, *1976* (R. Burlisch, R. D. Grigorieff, & J. Schröder, Eds). Lecture Notes in Mathematics 631. Berlin: Springer . pp. 188–200.

STETTER, H. J. 1979 Interpolation and error estimates in Adams PC-codes. *SIAM J. Numer. Anal.* **16,** 311–323.

STETTER, H. J. 1980 Tolerance proportionality in ODE-codes. In: *Proc. Second Conf. on Numerical Treatment of Ordinary Differential Equations* (R. März, Ed.) Seminarberichte No. 32. Berlin: Humboldt University. Pp. 109–123. Also in Working Papers for the 1979 SIGNUM Meeting on Numerical Ordinary Differential Equations, (R. D. Skeel, Ed.) Department of Computer Science, University of Illinois at Urbana-Champagne.

STEWART, N. F. 1970 Certain equivalent requirements of approximate solutions of $x' = f(t, x)$. *SIAM J. Numer. Anal.* **7,** 256–270.