Comments on the Third Lisbon V&V Workshop

Patrick J. Roache

The Lisbon III Workshop has been completed. Covering the full range of Code Verification, Calculation Verification, and Validation, the organizers and participants succeeded in maintaining the candor and camaraderie of the previous two Lisbon Workshops, while expanding the range and increasing the quality of the work. The ratio of discussion time to formal presentation time again approached unity, so that the Lisbon Workshops, unlike most "workshops," lived up to the name. Congratulations go to all involved.

In spite of all the good discussion, there are a few fundamental points that need re-stating or expanding. I would like to take this opportunity to do so.

Evaluation of Uncertainty Estimators for Small Samples

Several of the presentations and discussions described success, or at least promise of success, for new or modified $U_{95\%}$ estimators, based on a handful of test cases. The published literature contains methods justified on one or two cases, including the ITTC method. What could this mean? If we agree that the target is $U_{95\%}$, i.e. an estimator that includes the actual error in roughly 95% of the cases, then we need to examine at least 20 cases to expect to find one for which the actual error is outside the $U_{95\%}$ range. (Note this is not truly a "failure" because we do not want to pay the price of huge uncertainty estimates for ~100% coverage, including outliers.) Twenty cases would not be sufficiently large for dependable statistics; we would expect to need $O(100)$ cases to have confidence in the claim of roughly 95% coverage (as emphasized in [1]). "Cases" do not necessarily mean separate fluid dynamics problems, but include multiple grid triplets for the same problem, but for a convincing evaluation of a method for general use we certainly need many fluid dynamics problems.

In spite of this, I believe that the small sample studies are valuable. The analysts are essentially using their intuition, based on extensive computational experience, to cautiously infer something about the $U_{95\%}$ statistical coverage from the performance on a few cases. Suppose that we have just one case, and it is in some sense representative (i.e. not a singular problem, not a terrible grid, there are some points in the boundary layer and along the backstep, observed p is reasonable, etc.). If the $U_{95\%}$ estimate is 20% while the actual error is 0.1%, we would at least hypothesize (if not conclude) that the estimator is far too conservative, even though we all know that we cannot do any statistics with a sample size of one. Conceivably, we might do 100 total cases and find that it is non-conservative in the other 99 cases. But we are not doing blind statistics uninformed by experience, and if such a study showed 99/100 cases were non-conservative and 1/100 showed the estimator was far too conservative, we would surely re-examine that case, expecting to find that a mistake was made. Similarly, if we examined 5 cases and two were non-conservative but only by a little (say, $U_{95\%}$ estimator = 10% and actual error = 11%) we would be encouraged, even though if this pattern later proved to be replicated for 1000 cases, we would have verified not a $U_{95\%}$ estimator but an inadequate $U_{60\%}$ estimator.

Although these small sample studies, combined with well-founded intuition, are valuable, the convincing statistical evidence can come only with hundreds of cases, on some range of fluid dynamics problems. (Alternately, one could perform the evaluation only for a limited range of problems of interest, e.g. drag of ship hulls, claim only this success, and leave it to someone else to evaluate the $U_{95\%}$ estimator on other problems like airfoil lift, if that is their interest.)

With such statistically significant evaluations ultimately in mind, it seems to me ill-advised to be introducing too many new parameters into the $U_{95\%}$ estimators and fine tuning them on small sample studies. As I mentioned in the discussions, Dr. Jay Boris noted ironically in the 1980's that there were so many variations possible on discretization algorithms that nobody ever needed to use the same algorithm twice. The same is obviously true of RANS turbulence models, with all their switches and knobs, tunable for each new problem. Researchers are creative, and we are now in danger of seeing the same situation for uncertainty estimators. However, as our organizers Eça and Hoekstra have pointed out, it is also true that excessively conservative uncertainty estimates that "jump" by a factor 3/1.25 = 2.4 due to noise around observed p ~ 0.95 or 2.05 are disconcerting and would probably not be used. For me, this is an unsatisfying state of our present methodology, but at least it only occurs for the most difficult problems.

Exact Uncertainty?

What could "exact uncertainty" mean? In several presentations, when a $U_{95\%}$ estimator was evaluated, its value was said to be compared to the "exact uncertainty" or "true uncertainty." Actually, the authors (correctly) evaluated the $U_{95\%}$ estimator to see if it contained the "exact error" or "true error" in each case of a small-sample study. It is easy for any of us to slip up on terminology, and momentary confusion of *uncertainty* and *error* is common.

But there is another sense in which we could compare an uncertainty estimator with an "exact uncertainty" provided that the definition of uncertainty used is that used in V&V 20 [2]. Contrasted to the AIAA Guide [3] and to the ASME V&V 10 [4] (Computational Solid Mechanics), which define uncertainty in nebulous terms that cannot be implemented quantitatively, V&V 20 [2] defines numerical uncertainty as the standard deviation of the parent population. So if we had a population of (say) 100 different cases and we could meaningfully compare all with the same normalized metric (say drag coefficient) and for each case we had a numerical error, then the exact or true numerical uncertainty for that population of 100 cases would be calculated easily and unambiguously as the standard deviation of those numerical errors. (Technically, it is the sample standard deviation.)

Roughly 95%

If we predict a base drag coefficient to be 0.950 and an experiment gives 0.940, perhaps we would say that the prediction was close to acceptable. (Most people would say it was terrific.) But if we want *roughly* 95% coverage from an uncertainty estimator and we achieve "only" 94%, we do not say that it was merely close to acceptable. In fact, if it produced anywhere from 89% to 99% coverage, I would be more inclined to declare it a victory for this study and move on to

other fluid dynamics problems, rather than fine-tune parameters or introduce new *ad hoc* functions.

Richardson Extrapolation (RE) and the Asymptotic Range

We often heard during the Workshop that "Richardson Extrapolation only works in the asymptotic range", "RE only works for monotonic convergence", etc.

Only works for what? For error estimation, yes. But for use in uncertainty estimation, this is not true. The GCI based on RE (using either theoretical p, or observed over several grids) has been demonstrated to work (i.e. to provide roughly 95% uncertainty coverage) in hundreds of cases including many outside the asymptotic range and even with non-monotonic convergence. In the GCI, the absolute values around the RE error estimate and the Factor of Safety Fs >1 "cover a multitude of sins", to use an old expression. This is true even for Fs = 1.25, as in the work of the Terrassa group ([5]; see also [6]), which found GCI success for RANS turbulent axisymmetric flows even though as many as 1/3 of the nodes were converging non-monotonically. If, as I have recommended, we use Fs = 3 for cases of non-monotonic convergence, the original GCI uncertainty estimator is more robust. If one of the Least Squares variants of GCI is used, it is even more robust, especially with Fs = 3, as Eça and Hoekstra observe.

Dr. C. J. Freitas and his group at the Southwest Research Institute (U.S.A.) have applied the original GCI to many realistic problems with far from ideal convergence behaviors with good results. These include [7]: confined detonation (see V&V 20 [2], Section 2), vortex cavitation, blast-structure interactions, fluidized beds, shock propagation, structural failure, and turbulent multiphase flows, using FVM, FDM, FEM, ALE, and Riemann solvers. (Dr. Freitas intended to participate in this Workshop but his responsibilities and travel prevented it.)

$1/(r^p - 1)$, ε, and Data Range Δ_M

The fine-grid error estimate E from (generalized) Richardson Extrapolation RE is given by [8]

$$E_1 = \varepsilon / (r^p - 1)$$

where $\varepsilon = (f_2 - f_1)$ is the change in the quantity of interest over the last two grid calculations. The important functionality that the RE analysis displays is $1/(r^p - 1)$.

Several of the uncertainty estimators used at the Workshop utilize as a substitute for ε the data range Δ_M over a range of grids, somewhat subjectively defined as all the grids calculated, or those grids that seem to be in the acceptable range (also vaguely defined), or just the last two grids, in which case $\varepsilon = \Delta_M$. The data range is sometimes used as a substitute for E when the convergence behavior is poor (i.e. taking the uncertainty estimator as some $Fs \times \Delta_M$) and ε has been proposed as a candidate for the approximate error estimator itself. First, I observe that

perhaps only an *ordered* error estimator (with "efficiency factor" ~ 1 as defined by Hay and Pelletier in the Workshop) deserves to be called an error estimator in any practical engineering sense. Second, ε *is not an ordered error estimator* (nor is Δ_M). This was emphasized in the Workshop by Deng and Visonneau, which I appreciate. Note that both ε and Δ_M can be made (almost) arbitrarily small by choosing r close to unity, limited only by $r = 1 + 1/N$. My experience indicates that it is necessary to give a specific example to demonstrate just how serious this mistake can be. Consider a well behaved problem with $p = 1$, and a 100^D grid. Calculate the change in the solution ε for a change in the grid to 99^D. (I know that this is not practical for turbulent flows, but it can be done in principle, and in fact I have done it for Burgers equation in [8], page 124, and I want to make a point.) The Richardson Extrapolation shows that the ordered error estimate is

$$\varepsilon /[r^p - 1] = \varepsilon / [(100/99) - 1] = \varepsilon / [0.010101...] = 99 \times \varepsilon$$

So ε alone is unconservative by a factor of 99. Some estimator !

Since, for a given grid solution, we can make ε (almost) arbitrarily small *regardless of the accuracy* attained on that grid, how could ε deserve the dignity of the term "error estimator" or "approximate error"? Therefore neither ε nor Δ_M *generally* will work asymptotically as an error estimator, and we may expect that $Fs \times \varepsilon$ or $Fs \times \Delta_M$ *generally* will not work well asymptotically as an uncertainty estimator.

I want to be clear that I do not have a problem with using the data range Δ_M as a substitute of ε; for these problems, it is clearly an improvement. The problem I have is in taking either ε or the data range Δ_M to be an ordered error estimator by neglecting the term $1/(r^p - 1)$. One may protest that, for these anomalous behavior problems, we do not know (or trust) p. True, but at least we know r, and we are throwing away this knowledge if we substitute just Δ_M (or worse, just ε) for the ordered error estimate E. In fact, we are implicitly making an unjustified optimistic assumption about p. By taking the uncertainty estimator E to be $Fs \times \varepsilon$ or $Fs \times \Delta_M$, we are assuming $1/(r^p - 1) \sim 1$. This means we are assuming that $r^p \sim 2$. If our grid refinement factor $r \sim 1.4$ then we are effectively using $p \sim 2$. This is not good! If we do not know or trust p, perhaps we could at least assume $p = 1$, which would , for $r \sim 1.4$, introduce another scaling factor of $\sim 1 / 0.4 = 2.5$ compounded with whatever Fs we decide to use. Maybe even $p = 1$ is too optimistic, but it is surely more realistic than $p = 2$, giving a factor 2.5 vs 1 in the realistic case of $r \sim 1.4$. (Assuming $p = 0.5$ would be far too pessimistic and unpalatable for most users.) This neglected term is more important for smaller r, e.g. $r = 1.3$ minimum recommended by the new ASME JFE policy [9].

Then we also must ask, Why does it appear that $Fs \times \varepsilon$ works well as an uncertainty estimator in the Workshop paper of Hay and Pelletier using the solution-adaptive grid code CADYF?

Asymptotically, the ordered error estimator $E \sim \varepsilon$ only if $1/(r^p - 1) \sim 1$ or $r^p \sim 2$. For example, grid doubling with a first-order method gives $(r^p - 1) = 1$ and $E = \varepsilon$. Likewise for a second-order method with $r \sim 1.414$. For the finest two grids using CADYF, the adaptive grid

refinement algorithm produced near equidistribution of error estimates. Each subsequent grid refinement level is chosen to provide approximately a pre-set fractional reduction $\varsigma < 1$ in the error, and $\varsigma = [0.6, 0.8]$ are used. An assumed one-term power series expansion for discretization error shows that a reduction in E by a factor ς indicates $r^p = 1/\varsigma$ (regardless of the individual values of r and p) so that $\varsigma = 0.6$ produces $1/(r^p - 1) = 1.5$. Thus the $Fs = 3$ used by Hay and Pelletier is effectively $Fs = 4.5$ (for the last two adaptive grids) and is quite conservative, even more so for $\varsigma = 0.8$ (which gives effective $Fs = 12$). If the actual error reduction was somewhat better than the nominal or targeted $\varsigma = 0.6$, it would make the process less conservative; $\varsigma = 0.5$ would produce $E = \varepsilon$. So the good performance of the CADYF uncertainty estimator is not due to the fact that $\varepsilon \sim E$ *in general*, but only for the particulars of the refinement, as recognized by the authors (end of their section 3). In any case, it is a pleasure to see the *Fs* approach successfully applied to the ZZ and Wiberg error estimators to obtain uncertainty estimates.

Although oscillatory convergence obviates the use of direct (non-Least-Squares) calculations of p from successive grid triplets, and contradicts the assumptions of the one-term power series representation of discretization error, it seems intuitive that the term $1/(r^p - 1)$ might still represent the convergence scaling of the envelope of the oscillatory convergence, and could be applied to Δ_M in the GCI. It would be interesting to examine this premise using a two-term power series representation, since Roy [10] has shown that this can explain one cause of oscillatory convergence for mixed-order discretizations (i.e. the common first-order advection and second-order diffusion). In any case, we know it works, very often, in empirical tests. And in the limit, if convergence becomes monotonic, it certainly provides an ordered error estimate E, which ε or Δ_M alone does not. (This also applies to the AES estimator of Celik and colleagues (Elizalde-Blancas et al.) in this Workshop; but their use of a global "proportionality constant c", which takes the place of $1/(r^p - 1)$, gives some improvement.) Even if ε or Δ_M alone "works" with some *Fs* for some grid range, surely there is no inherent value in *not* being an ordered error estimator, as Deng and Visonneau would agree. RE works for error estimate E in the best cases; there is no advantage in replacing it with an *ad hoc* prescription that does not work even in the best cases.

My original GCI paper emphasized the importance of the scaling contained in the term $1/(r^p - 1)$. Rather than 95% coverage, the original target of the scaling was to re-normalize the ε obtained from studies with various r and p for comparison to the ε that would be obtained from a grid doubling with a second-order method, i.e. $r = 2$, $p = 2$. (The title of the paper was "A Method for Uniform Reporting of Grid Refinement Studies." See [8].) My own elicitation of expert opinion from $O(200)$ CFD practitioners had indicated that ε from a grid doubling with a second-order method provided a practical level of confidence. (Only after analysis was it shown that this corresponded to $Fs = 3$, which corroborated the expert opinion.) For example, without any quantitative claims for 95% coverage or 99% coverage, how would we compare two studies of the same problem if one study used $r = 1.5$, $p = 1$ and produced $\varepsilon = 3\%$, and another study used $r = 2$, $p = 2$ and produced $\varepsilon = 6\%$? Which solution would be more dependable? The scaling alone (without deciding on *Fs*) shows that the ε for the first case should be scaled to 18% to compare with the 6% of the second case. Whatever level of coverage is intended, and whatever *Fs* is used, the $1/(r^p - 1)$ scaling is critical for the reporting**.**

Error estimation from a grid triplet without explicit evaluation of *p*

We would all like to avoid assuming a power series for the convergence, or using Taylor series, in order to be more general. But it seems to me that any general method which one hopes to apply to difficult cases ought to work on the well behaved cases first. When the problem is well behaved, the Taylor series or power series works well.

The AES method of Elizalde-Blancas et al. in this Workshop utilizes the idea that the extrapolation to f_{exact} can be accomplished (in some cases) without explicitly evaluating observed p. Following their approach, but assuming a well behaved problem with constant r and p, the procedure for Richardson error estimate E_1 leads to an interesting result. E_1 for the fine grid solution f_1 and E_2 for the medium grid solution f_2 are taken from page 112 of [8], using the dimensional form of Eqs. 5.4.2 and 5.4.3, as

(A) $E_1 \equiv f_1 - f_{exact} = [f_2 - f_1] / [r^p - 1]$

(B) $E_2 \equiv f_2 - f_{exact} = [f_3 - f_2] / [r^p - 1]$

Subtracting (A) from (B) leads to

(C) $1/[r^p - 1] = [f_2 - f_1] / [f_3 - 2f_2 + f_1]$

Comparing this to Elizalde-Blancas et al. shows it is equivalent to their Eq. 4 with their

(D) $C = 1/[r^p - 1]$.

(It is not equivalent to their Eq. 5.) Eq. C leads to a nice formula for the error estimation that does not involve r or p (although this could be misleading, since constant r and p are necessary requirements). Using C in A gives

(E) $E_1 = [f_2 - f_1]^2 / [f_3 - 2f_2 + f_1]$

Using the ε notation (Eq. 5.4.3 of [8], dimensional) with

(F) $\varepsilon_{21} = f_2 - f_1$ and $\varepsilon_{32} = f_3 - f_2$

then E becomes

(H) $E_1 = \varepsilon_{21}^2 / [\varepsilon_{32} - \varepsilon_{21}] = \varepsilon_{21} / [(\varepsilon_{32} / \varepsilon_{21}) - 1]$

The latter form is better behaved for round-off error, and allows a confidence check to see if the results are consistent with expected convergence behavior. In the usual method, the observed p is compared to the theoretical p. As stressed by Hoekstra at the Workshop, agreement for a single grid triplet is no guarantee of monotone behavior, but it is comforting, and more confidence is

built with more grid triplets. In the present approach, we would look for $[(\varepsilon_{32} / \varepsilon_{21}) - 1] \sim [r^p - 1]$ using known r and expected (theoretical) p.

I have tested Eq. H on synthetic problems, for which it is exact, and for practical problems given in [8] on pages 164, 177, 181. For the very well behaved first problem on page 177, the method gives a factor of 2.956 when the assumed $r = 2$ and $p = 2$ would give a factor of 3. The GCI calculation with $p = 2$ shows that the GCI ratio (for $r = 2$), which should = 4 for these assumptions, is actually 3.95. So both the GCI calculation with explicit evaluation of p, and Eq. H, are consistent with observed p just slightly less than 2. For the airfoil problems on page 164 (turbulent flow) and page 181 (inviscid), the convergence behavior is noticeably not ideal but fairly well behaved, and the p - free extrapolation seems to work as well as the usual method with explicit evaluation of p.

Quality Validation Experiments

Quality validation experiments are rare. The backstep problem of Driver and Seegmiller [11] used in the Workshop is often cited as an example of high quality experimental work, yet we discovered upon careful reading that there was no measurement of inflow boundary layer properties at the upper wall, and no experimental uncertainty given for the time-averaged velocity components or Reynolds stresses. When the V&V 20 [2] project was begun, it was intended to use real experimental data on a heat exchanger for the example end-to-end problem of the document. No adequate experiment, with thorough measurement of inflow and outflow values and uncertainties, could be found. Instead, synthetic experimental values were used to illustrate the V&V procedures. The follow-on document to V&V 10 [4] (Computational Solid Mechanics) is now in development. The original intention was to use real experimental data for a simple structural problem, but no adequate experiment could be found; instead, synthetic experimental values will be used. The ASCE V&V document for free surface flow V&V [12] has plenty of real experimental data, ranging from laboratory scale to site scale (e.g. sections of the Mississippi River, San Francisco Bay), but uncertainty statements are rare. In a 5-day Workshop on Combustion Modeling [13] held in Pittsburgh in February 2006, only one experimental paper (on stationary gas turbines at NSF) had complete description of inflow properties - not uncertainties, just measurements of the inflow properties. In a 5-day workshop [14] on Nuclear Reactor modeling held in Idaho in July 2008, only one paper had any experimental uncertainty estimates; this work by Dr. H. McIlroy and colleagues [15] is an exceptionally high quality PIV (particle image velocimetry) experiment designed expressly for CFD validation, with complete field measurements (including inflow) and complete uncertainty measurements. It is so rare as to be perhaps singular for these flows. The benchmark study for validation experiments in aerodynamics remains Aeschliman and Oberkampf [16]. The great majority of experimental works not only fail to report experimental uncertainties but fail to measure all pertinent inflow quantities, so that results could not be compared with other *experiments* - forget CFD. Uncontrolled, unmeasured experiments are the norm, and experimental work is the weakest link in V&V.

U_{val} vs. Error Bars

Only two of the Workshop papers actually calculated and interpreted U_{val} following the abbreviated V&V 20 procedure as described in the Workshop announcement. The interpretation this enables is an improvement over just plotting computations with their uncertainties and experiments with their uncertainties, as done in the remaining papers. However, just this "error bar" approach is a huge improvement over typical CFD validation reporting. Both the U_{val} and the error bar presentations are valuable for interpreting model fidelity, and in fact, if I could have only one, I would want the error bar presentation. But it is little trouble to produce both, and I hope to see more acceptance of this approach among Workshop participants and their colleagues, as well as the general CFD community.

Final Comment

The Lisbon III Workshop was a worthy finale to the series. When I told a colleague, a well-known experimentalist, that the ratio of discussion time to presentation time approached unity, he gasped in disbelief and envy. Equally important is the immeasurable quality of the interactions. I feel privileged to have participated in all three Lisbon V&V Workshops, and I cherish the resulting relationships. My congratulations and gratitude go to our organizers, Luis Eça and Martin Hoekstra.

References

1. Roache, P. J. (2003), "Error Bars for CFD," AIAA Paper 2003-0408, AIAA 41st Aerospace Sciences Meeting, January 2003, Reno, Nevada.
2. ASME Committee PTC-61, *ANSI Standard V&V 20. ASME Guide on Verification and Validation in Computational Fluid Dynamics and Heat Transfer*, (expected) 2008.
3. AIAA, 1998, *Guide for the Verification and Validation of Computational Fluid Dynamics Simulations*, AIAA G-077-1998, American Institute of Aeronautics and Astronautics, Reston, VA.
4. ASME Committee PTC-60, *ANSI Standard V&V 10. ASME Guide on Verification and Validation in Computational Solid Mechanics*, 29 December 2006.
5. Cadafalch, J., Pérez-Segarra, C. C., Cónsul, R. and Oliva, A., "Verification of Finite Volume Computations on Steady State Fluid Flow and Heat Transfer", ASME Journal of Fluids Engineering, Vol. 124, pp. 11-21, March 2002.
6. Roache, P. J. (2003), "Conservatism of the GCI in Finite Volume Computations on Steady State Fluid Flow and Heat Transfer", *ASME Journal of Fluids Engineering*, Vol. 125, No. 4, July 2003, pp 731-732.
7. Freitas, C. J. (2008), pers. comm.
8. Roache, P. J. (1998), *Verification and Validation in Computational Science and Engineering*, Hermosa Publishers, New Mexico, 1998.
9. Celik, Ismail. B., Ghia, Urmila, Roache, Patrick J. and Freitas, Christopher J. (2008), "Procedure for Estimation and Reporting of Uncertainty due to Discretization in CFD Applications", *ASME Journal of Fluids Engineering*, Editorial.
10. Roy, C. J. (2003), "Grid Convergence Error Analysis for Mixed-Order Numerical Schemes," *AIAA Journal*, Vol. 41, No. 4, April 2003, pp. 595-604. See also AIAA Paper 2001-2606, Anaheim, June 2001.
11. Driver, D. M. and Seegmiller, H. L. (1985), "Features of a Reattaching Turbulent Shear layer in Divergent Channel Flow," *AIAA Journal*, Vol. 23, No. 2, February 1985, pp. 163-171.
12. ASCE/EWRI Task Committee on 3D Free Surface Flow Model Verification/Validation, *3D Free Surface Flow Model Verification/Validation*, *2009.*
13. *Proc. Workshop on the Future of Modeling and Simulation for Combustion Applications*, Pittsburgh, PA, Feb. 21-23, 2006.

14. *Proc. Verification and Validation for Nuclear Systems Analysis Workshop, Idaho National Laboratory,* Idaho Falls, Idaho, 21-25 July 2008.
15. McIlroy, H. M. Jr., McEligot, D. M. and Pink, R. J., 2007, "Measurement of Turbulent Flow Phenomena for the Lower Plenum of a Prismatic Gas-Cooled Reactor," Paper ID 067, Proceedings of the 12th International Topical Meeting on Nuclear Reactor Thermal Hydraulics (NURETH-12), Pittsburgh, PA, September 30 - October 4, 2007. See also [14].
16. Aeschliman, D. P., and Oberkampf, W. L., "Experimental Methodology for Computational Fluid Dynamics Code Validation," *AIAA Journal*, Vol. 36, No. 5, 1998, pp. 733-741.