STEADY STATE MODELS and the INITIAL TRANSIENT

There are five main methods for dealing with an initial transient:
1. Run-in model for a warm-up period till reaches steady state. Delete data from the warm-up period.
2. Set initial conditions of the model so that model starts in steady-state.
3. Set partial initial conditions then warm-up model and delete warm-up data.
4. Run model for very long time making the bias effect negligible.
5. Estimate steady state parameters from a short transient simulation run.

This project is only concerned with the first method: Deletion of the initial transient data by specifying a warm-up period.

Question is: How do you estimate the length of the warm-up period required?

There are five main methods:
1. GRAPHICAL METHODS: Involve visual inspection of the time-series output and human judgement.
2. HEURISTIC APPROACHES – Provide rules for determining when initialisation bias has been removed.
3. STATISTICAL METHODS – Based upon statistical principles.
4. INITIALISATION BIAS TESTS – These test whether the warm-up period has been deleted. They test the null hypothesis, H_0: no initialisation bias present in a series of data.
5. HYBRID METHODS – Combines initialisation bias tests with graphical or heuristic methods to determine warm-up period.

This document contains details of 42 such methods that were found by the authors of this document in published literature. Each method has been given a unique ID number, as has each journal/book reference relating to the methods. The first summary table (Table1) gives the method ID, method name, method type and all journal/book ref IDs relating to that method, for each of the 42 methods found. The second summary table (Table2) gives the reference ID, author names, title and journal/book information for each of the papers/books referenced. The following section, titled "Warm up Methods: Summary Details and Literature Review**",** contains individual tables, sorted into method types, which give the details of each warm up method. Each table consists of the method ID and name, a brief description of the method, a brief literature review and a summary of the positives and negatives (criticisms) of the method. Each method is also rated on accuracy, simplicity, automation potential (particularly important for this project), generality, including the number of parameters requiring estimation and computation time.

Warmup Methods			
Method ID	**Method Name**	**Method Type**	**Paper Refs ID**
1	Simple Time Series Inspection	Graphical	48
2	Ensemble (Batch) Average Plots	Graphical	51
3	Cumulative-Mean Rule	Graphical	48, 35, 32, 16, 57, 6, 51, 37, 4, 45
4	Deleting-The-Cumulative-Mean Rule	Graphical	57, 6
5	CUSUM Plots	Graphical	16
6	Welch's Method	Graphical	34, 7, 53, 52, 51, 36, 4, 1, 45
7	Variance Plots (or Gordon Rule)	Graphical	48, 35, 32, 7
8	Statistical Process Control Method (SPC)	Graphical	52, 1, 8
9	Ensemble (Batch) Average Plots with Schribner's Rule	Heuristic	35, 13, 7
10	Conway Rule or Forward Data-Interval Rule	Heuristic	40, 13, 32, 35, 50, 7, 23, 5, 1, 49
11	Modified Conway Rule or Backward Data-Interval Rule	Heuristic	35, 32, 5, 58
12	Crossing-Of-The-Mean Rule	Heuristic	35, 32, 13, 7, 5, 58, 1
13	Autocorrelation Estimator Rule	Heuristic	24, 35, 7
14	Marginal Confidence Rule or Marginal Standard Error Rules (MSER)	Heuristic	5, 28, 36
15	Marginal Standard Error Rule m, (e.g. m=5, MSER-5)	Heuristic	28, 1, 45
16	Goodness-Of-Fit Test	Statistical	7
17	Relaxation Heuristics	Heuristic	46, 7, 57, 6, 36
18	Kelton and Law Regression Method	Statistical	11, 34, 46, 7, 57, 6, 47, 52, 36
19	Randomisation Tests For Initialisation Bias	Statistical	23, 1
20	Schruben's Maximum Test (STS)	Initialisation Bias Tests	20, 34, 18, 23, 22, 52
21	Schruben's Modified Test	Initialisation Bias Tests	16, 34, 28, 52

Warmup Methods			
Method ID	**Method Name**	**Method Type**	**Paper Refs ID**
22	Optimal Test (Brownian bridge process)	Initialisation Bias Tests	18, 46, 7, 31, 52
23	Rank Test	Initialisation Bias Tests	14, 31, 52
24	Batch Means Based Tests - Max Test	Initialisation Bias Tests	33, 42, 43, 52, 28
25	Batch Means Based Tests - Batch Means Test	Initialisation Bias Tests	33, 43, 22, 28, 52
26	Batch Means Based Tests - Area Test	Initialisation Bias Tests	33, 43, 22, 52
27	Pawlikowski's Sequential Method	Hybrid	7
28	Scale Invariant Truncation Point Method (SIT)	Hybrid	17
29	Exponentially Weighted Moving Average Control Charts	Graphical	9
30	Algorithm for a Static Dataset (ASD)	Statistical	4
31	Algorithm for a Dynamic Dataset (ADD)	Statistical	4
34	Telephone Network Rule	Heuristic	19
35	Ockerman & Goldsman Students t-tests Method	Initialisation Bias Tests	22
36	Ockerman & Goldsman (t-test) Compound Tests	Initialisation Bias Tests	22
37	Glynn & Iglehart Bias Deletion Rule	Statistical	38
38	Wavelet-based spectral method (WASSP)	Statistical	39, 54, 56
39	Queueing approximations method (MSEASVT)	Statistical	41
40	Chaos Theory Methods (methods M1 and M2)	Statistical	42
41	Beck's Approach for Cyclic output	Heuristic	44
42	Tocher's Cycle Rule	Heuristic	7
43	Kimbler's Double exponential smoothing method	Heuristic	46
44	Kalman Filter method	Statistical	47, 52
45	Euclidean Distance (ED) Method	Heuristic	58

Warmup Methods			
Method ID	**Method Name**	**Method Type**	**Paper Refs ID**
46	Neural Networks (NN) Method	Heuristic	58

Table 1: Summary of all warm-up methods found in literature: method ID, method name, method type and all journal/book ref IDs relating to each method.

(Return to top of document)

References										
Ref ID	**Authors**	**Title**	**Journal**	**Year**	**Volume**	**Number**	**Pages**	**Publisher**	**Type**	**Book?**
1	Mahajan, P. S., and Ingalls, R. G.	Evaluation of methods used to detect warm-up period in steady state simulation	Proceedings of the Winter Simulation Conference	2004			663-671		Comparison only	No
4	Bause, F., and Eickhoff, M.	Truncation point estimation using multiple replications in parallel	Proceedings of the Winter Simulation Conference	2003			414-421		New Method	No
5	Preston White Jnr, K.	An effective truncation heuristic for bias reduction in simulation output	Simulation	1997	69	6	323-334		New Method	No
6	Roth, E	The relaxation time heuristic for the initial transient problem in M/M/k queueing systems	European Journal of Operational Research	1994	72		376-386		New Method	No
7	Pawlikowski, K	Steady-state simulation of queueing processes: A survey of problems and solutions	Computing Surveys	1990	122	2	123-170		Survey	No
8	Robinson, S	A statistical process control approach to selecting a warm-up period for a discrete-event simulation	European Journal of Operational Research	2005	176		332-346		New Method	No

References										
Ref ID	**Authors**	**Title**	**Journal**	**Year**	**Volume**	**Number**	**Pages**	**Publisher**	**Type**	**Book?**
9	Rossetti, M. D., Li, Z., and Qu, P	Exploring exponentially weighted moving average control charts to determine the warm-up period	Proceedings of the Winter Simulation Conference	2005			771-780		New Method	No
11	Kelton, W. D., and Law, A. M	A new approach for dealing with the startup problem in discrete event simulation	Naval Research Logistics Quarterly	1983	30		641-658		New Method	No
13	Wilson, J. R., and Pritsker, A. A. B.	Evaluation of startup policies in simulation experiments	Simulation	1978	31	3	79-89		Comparison only	No
14	Vassilacopoulos, G.	Testing for initialization bias in simulation output	Simulation	1989	52	4	151-153		New Method	No
16	Nelson, B. L	Statistical analysis of simulation results	Handbook of industrial engineering	1992				John Wiley NY, 2nd Ed	Survey	Yes
17	Jackway, P. T., and deSilva, B. M.	A methodology for initialisation bias reduction in computer simulation output	Asia-Pacific Journal of Operational Research	1992	9		87-100		New Method	No
18	Schruben, L., Singh, H., and Tierney, L.	Optimal tests for initialization bias in simulation output	Operations Research	1983	31	6	1167-1178		New Method	No
19	Zobel, C. W., and Preston White Jnr, K.	Determining a warm-up period for a telephone network routing simulation	Proceedings of the Winter Simulation Conference	1999			662-665		New Method	No
20	Schruben, L. W.	Detecting initialization bias in simulation output	Operations Research	1982	30	3	569-590		New Method	No
22	Ockerman, D. H., and Goldsman, D	Student t-tests and compound tests to detect transients in simulated time series	European Journal of Operational Research	1999	116		681-691		New Method	No
23	Yucesan, E	Randomization tests for initialization bias in simulation output	Naval Research Logistics	1993	40		643-663		New Method	No
24	Fishman, G. S.	Estimating sample size in	Management Science	1971	18		21-38		New Method	No

References										
Ref ID	**Authors**	**Title**	**Journal**	**Year**	**Volume**	**Number**	**Pages**	**Publisher**	**Type**	**Book?**
		computing simulation experiments								
28	Preston White Jnr, K., Cobb, M. J., and Spratt, S. C	A comparison of five steady-state truncation heuristics for simulation	Proceedings of the Winter Simulation Conference	2000			755-760		Comparison only	No
31	Ma, X., and Kochhar, A. K.	A comparison study of two tests for detecting initialization bias in simulation output	Simulation	1993	61	2	94-101		Comparison only	No
32	Gafarian, A. V., Ancker Jnr, C. J., and Morisaku, T.	Evaluation of commonly used rules for detecting "steady state" in computer simulation	Naval Research Logistics Quarterly	1978	25		511-529		Comparison only	No
33	Cash, C. R., Dippold, D. G., Long, J. M., and Pollard, W. P.	Evaluation of tests for initial-condition bias	Proceedings of the Winter Simulation Conference	1992			577-585		Comparison only	No
34	Law, A. M.	Statistical analysis of simulation output data	Operations Research	1983	31		983-1029		Survey	No
35	Wilson, J. R., and Pritsker, A. A. B	A survey of research on the simulation startup problem	Simulation	1978			55-58		Comparison only	No
36	Linton, J. R., and Harmonosky, C. M.	A comparison of selective initialization bias elimination methods	Proceedings of the Winter Simulation Conference	2002			1951-1957		Comparison only	No
37	Fishman, G. S.	Discrete-Event Simulation		2001				Springer	Instructional incl. warmup	Yes
38	Glynn P W and Iglehart D L	A New Initial Bias Deletion rule	Proceedings of the Winter Simulation Conference	1987			318-319		New Method	No
39	Lada E K, Wilson J R and Steiger N M.	A wavelet-based spectral method for steady-state simulation analysis	Proceedings of the Winter Simulation Conference	2003			422-430		New Method	No

References										
Ref ID	**Authors**	**Title**	**Journal**	**Year**	**Volume**	**Number**	**Pages**	**Publisher**	**Type**	**Book?**
40	Conway R W	Some tactical problems in digital simulation	Management Science	1963	10	1	47-61		New Method	No
41	Rossetti M D and Delaney CPT P J	Control of initialization bias in queueing simulations using queueing approximations	Proceedings of the Winter Simulation Conference	1995			322-329		New Method	No
42	Lee Y-H and Oh H-S	Detecting truncation point in steady-state simulation using chaos theory	Proceedings of the Winter Simulation Conference	1994			353-360		New Method	No
43	Goldsman D, Schruben L W and Swain J J	Tests for transient means in simulated time series	Naval Research Logistics	1994	41		171-187		New Method	No
44	Beck A D	Consistency of warm up periods for a simulation model that is cyclic in nature	Proceedings of the Simulation study group (OR Society)	2004			105-108		New Method	No
45	Sandikci B and Sabuncuoglu I	Analysis of the behaviour of the transient period in non-terminating simulations	European Journal of Operational Research	2006	173		252-267		Comparison only	No
46	Kimbler D L and Knight B D	A survey of current methods for the elimination of initialisation bias in digital simulation	Annual Simulation Symposium	1987	20		133-142		Comparison only	No
47	Gallagher M A, Bauer K W Jnr and Maybeck P S	Initial Data truncation for univariate output of discrete-event simulations using the Kalman Filter	Management Science	1996	42	4	559-575		New Method	No
48	Gordon G	System simulation		1969				Prentice-Hall NJ	New Method	Yes
49	Fishman G S	Concepts and methods in discrete event digital simulation		1973				Wiley NY	New Method	Yes
50	Bratley P, Fox B and Schrage L	A guide to simulation (2nd Ed)		1987				Springer-Verlag NY	Survey	Yes

References										
Ref ID	**Authors**	**Title**	**Journal**	**Year**	**Volume**	**Number**	**Pages**	**Publisher**	**Type**	**Book?**
51	Banks J, Carson J S, Nelson, B L and Nicol D M	Discrete-event system simulation		2001				Prentice Hall NJ	Survey	Yes
52	Law A M and Kelton W D	Simulation Modelling and Analysis		2000				McGraw-Hill NY	Survey	Yes
53	Alexopoulos C and Seila A F	Output data analysis	Handbook of Simulation (Ed: J Banks)	1998			225-272	Wiley NY	Survey	Yes
54	Lada E K, Wilson J R, Steiger N M and Joines J A	Performance evaluation of a wavelet-based spectral method for steady-state simulation analysis	Proceedings of the Winter Simulation Conference	2004			694-702		New Method	No
56	Lada E K and Wilson J R	A wavelet-based spectral procedure for steady-state simulation analysis	European Journal of Operational Research	2006	174		1769-1801		New Method	No
57	Roth E and Josephy N	A relaxation time heuristic for exponential-Erlang queueing systems	Computers & Operations research	1993	20	3	293-301		New Method	No
58	Lee Y-H, Kyung K-H and Jung C-S	On-Line determination of steady state in simulation outputs	Computers industrial engineering	1997	33	3	805-808		New Method	No

Table 2: Summary of the literature searched for warm-up methods: reference ID, author names, title and journal/book information.

(Return to top of document)

WARM-UP METHODS: SUMMARY DETAILS and LITERATURE REVIEW

CRITERIA	RATINGS	PREFERRED RATING
Accuracy	Good/Medium/Poor	Good
Simplicity	Good/Medium/Poor	Good/Medium
Automation potential	Good/Medium/Poor	Good/Medium
Generality (No assumptions)	Good/Medium/Poor	Good/Medium
Parameters to estimate?	None,1, 2, etc… (V) = (incl. variance)	None/Few
Computation time	Good/Medium/Poor	Good/Medium

(Return to top of document)

HEURISTIC METHODS

Method ID	9
Method Name	Ensemble (Batch) Average Plots with Schribner's Rule
Brief description	Pawlikowski (1990) [7]: "In a time series of batch means $\bar{X}_1(m_0), \bar{X}_2(m_0),...,$ the initial transient is over after b_0 batches, that is, after $n_0 = b_0 m_0$ observations, if the k most recent batch means all fall within an interval of width δ_1; that is if: $\left|\bar{X}_{b_0-i}(m_0) - \bar{X}_{b_0-j}(m_0)\right| < \delta_1$, for 0<i<k-1, 0<j<k-1
Literature Review	Pawlikowski (1990) [7]: Method "…is sensitive to the value of the parameter k [number of batch means which must fall within interval], which should depend on the variability of the observed process" Wilson & Pritsker (1978b)[13]: Test Method: Used a 5 step procedure for evaluating 3 alternative start-up policies with 4 different truncation methods:

	IC1: initial conditions are idle & empty, IC2: start as close to steady state mode as possible, IC3: start as close to steady state mean as possible; Truncation Rules: TR0: retain all data, TR1: Use Conway rule (10), TR2: Use Crossing-of-the-mean rule (12), TR3: Use Ensemble (Batch) Average Plots with Schribner's Rule (9): Step1: Select standard system and compute bias, variance and MSE for all possible values of truncation point and for each initial condition (ICi) as specified above. Step2: Estimate distribution of truncation pt over independent simulation runs for each start up policy (ICi, TRi). Step3: Compute average bias, var & MSE by combining results from Steps1 & 2. Step4: Select a base policy & for every truncation pt under this policy compute the 1/2 length of a confidence interval (CI) for μ centred at the sample mean for that truncation pt. Average these values with respect to the observed distribution from Step2 to create a standard average CI 1/2 length for the base policy. Adjust CIs for all other policies so their average 1/2 length = the base average 1/2 length. Step5: Calculate probabilities that a CI centred around the estimated mean will actually cover the true mean. Calculate these coverage probabiities for all possible values of truncation pt, and average results w.r.t the appropriate truncation pt distribution for each policy. Systems Studied: M/M/1 queue with a capacity of 15 and a traffic intensity (ρ) of 0.9; A Machine Repair System - each machine in a group of 14 independent units has a constant failure rate of 0.2 so that time to failure is exponentially distributed with a mean of 5. Repair station rule is first come first served, with 3 parallel servers & enough waiting room for all machines. Repair times are exponentially distributed with a mean of 2. Arrival rate depends on how many units are in the system. Results: Ensemble (Batch) Average Plots with Schribner's Rule (9) & Crossing-of-the-mean rule (12) seemed to reduce the bias much more than the Conway rule (10) but suffered from a loss of CI coverage regarding the true mean. This showed apparently that a misspecification of the parameter/s can cause a large loss of coverage. Wilson & Pritsker (1978a) [35]: Requires 3 parameters to be specified: Batch size, batch count & tolerance.
Positives	Simple. Automatable.
Criticisms	Requires 3 parameters to be specified: Batch size, batch count & tolerance [35]. Misspecification of the

	parameter/s can cause a large loss of coverage [13]. Sensitive to the value of the parameter, k, number of batch means which must fall within interval, which should depend on the variability of the observed process. [7].
Accuracy	Medium/Poor
Simplicity	Good
Automation potential	Good
Generality (No assumptions)	Good
Parameters to estimate?	3 - Batch size, batch count (k) & tolerance (δ)
Computation time	Good

Method ID	10
Method Name	Conway Rule (or Forward Data-Interval Rule)
Brief description	Make pilot runs. Collect observations at relatively short intervals. Bratley et al (1987) [50]: For each pilot run, "…discard initial observations until the first one left is neither the maximum nor the minimum of the remaining observations." Fishman (1973) [49]: Choose the largest truncation point out of all the pilot runs. Recommended that it be applied to batch means and not original observations. Conway [40] cautioned "against examining cumulative statistics …as such will typically lag behind the current state."
Literature Review	Mahajan & Ingalls (2004) [1]: Systems Studied: Job Shop model consisting of 5 Cells, C(i), i=1…5. Each cell has different number of machines (resources). There are 3 customer classes A,B,C. Overall arrival rate is poisson. Service times are exponential with mean dependent on customer class and cells. Arriving parts are split into classes {A, B, C} with probability {0.5, 0.3, 0.2}. 3 types of utilisation are used: TypeI is high utilisation with an average utilisation of 90% and range 80-95%; TypeII is moderate with an average utilisation of 70% and range 65-80%; TypeIII is Low with average utilisation of 50% with range 45-65%. Models are started empty & idle. Initial run length is 1000 hrs, with 5 replications. Performance criteria: Final MSE & Variance, average computing time, percentage change in MSE, percentage change in variance. Method said to perform well if it reduces both MSE and Variance and is

	computationally efficient. Results: Method performed well with the low utilised systems and is therefore recommended for use with those types of systems. Preston White Jnr (1997) [5]: Systems Studied: Four M/M/1 queue systems with output variable set as number in system: 1. Empty Queue: model starts empty & idle with ρ = 0.9; 2. Loaded Queue: 100 entities in system at start with ρ = 0.9; 3. Filling Tandem Queue: model starts empty & idle with ρ = 1.4 and a capacity of 100; 4. Transient Tandem Queue: model starts empty & idle with ρ = 1.4; Tested 4 methods: Conway rule (10), Modified Conway rule (11), Crossing-of-the-mean rule (12) and MCR (14). Results: All better than not truncating at all! "For the empty queue experiments, all of the rules give reasonably accurate overall point estimates for the grand mean (within 3% of population mean), with the sole exception of the un-weighted Modified Conway rule (11). Only MCR (14) provides coverage of the 95% confidence interval however." "The results for .. [no truncation of data] and [Conway rule (10)] are identical to each other…and also provide the most precise interval estimate as expected. The results differ only on one run [out of 10] and provide the most accurate point estimate." "For the loaded queue experiments, …all of the rules provide significant correction [of bias] through truncation. Again only the MCR (14) yields coverage of the 95% CI." "For the filling queue experiments, …the rules provide correction through truncation, with remarkably consistent point and interval estimates of the grand sample mean across rules and across runs." Pawlikowski (1990) [7]: "...poor approximation of the duration of the initial transient. As was shown in Gafarian et al. (1978) [32], using this rule we can significantly overestimate the length of the initial transient for small ρ and underestimate it for high ρ; (see also Wilson and Pritsker 1978b [13]). " Wilson & Pritsker (1978b) [13]: Test Method: Used a 5 step procedure for evaluating 3 alternative start-up policies with 4 different truncation methods:

	IC1: initial conditions are idle & empty, IC2: start as close to steady state mode as possible, IC3: start as close to steady state mean as possible; Truncation Rules: TR0: retain all data, TR1: Use Conway rule (10), TR2: Use Crossing-of-the-mean rule (12), TR3: Use Ensemble (Batch) Average Plots with Schribner's Rule (9): Step1: Select standard system and compute bias, variance and MSE for all possible values of truncation point and for each initial condition (ICi) as specified above. Step2: Estimate distribution of truncation pt over independent simulation runs for each start up policy (ICi, TRi). Step3: Compute average bias, var & MSE by combining results from Steps1 & 2. Step4: Select a base policy & for every truncation pt under this policy compute the 1/2 length of a confidence interval (CI) for μ centred at the sample mean for that truncation pt. Average these values with respect to the observed distribution from Step2 to create a standard average CI 1/2 length for the base policy. Adjust CIs for all other policies so their average 1/2 length = the base average 1/2 length. Step5: Calculate probabilities that a CI centred around the estimated mean will actually cover the true mean. Calculate these coverage probabiities for all possible values of truncation pt, and average results w.r.t the appropriate truncation pt distribution for each policy. Systems Studied: M/M/1 queue with a capacity of 15 and a traffic intensity (ρ) of 0.9; A Machine Repair System - each machine in a group of 14 independent units has a constant failure rate of 0.2 so that time to failure is exponentially distributed with a mean of 5. Repair station rule is first come first served, with 3 parallel servers & enough waiting room for all machines. Repair times are exponentially distributed with a mean of 2. Arrival rate depends on how many units are in the system. Results: Ensemble (Batch) Average Plots with Schribner's Rule (9) & Crossing-of-the-mean rule (12) seemed to reduce the bias much more than the Conway rule (10) but suffered from a loss of CI coverage regarding the true mean. Yucesan (1993) [23]: Randomisation Test (19) generally more conservative than Conway rule (10) or (graphical) Schruben (20). Gafarian et al. (1978) [32]: Systems Studied: M/M/1 queue with output variable as waiting time in queue per customer; ρ = 0.1, 0.5,

	0.7, 0.9. Method is tested 100 times. 'Goodness' (Performance) Criteria: (i) Accuracy: The ratio, mean of [estimated truncation values] / [true truncation value], is calculated and a value near 1 implies method accurate. (ii) Precision: Measure of variation in estimated truncation value is calculated as Sqrt(variance of estimated truncation points) / (mean of estimated truncation points). A value close to zero implies method precise. (iii) Generality: Judged to be general if the rule performs well across a broad range of systems and parameters within a system. (iv) Cost: Expense in computer time. (v) Simplicity: Accessible to average practitioner. (i) Accuaracy, (ii) Precision and (iii) Generality are considered first. Any method not satisfactory on all three is discarded. Computer cost is a last priority. Results: Conway rule (10) satisfied the simplicity criteria but failed the accuracy criteria. It overestimated for low ρ, grossly underestimated for high ρ and failed to find a truncation pt on many occasions (especially when run length was small). Wilson & Pritsker(1978) [35]: Comments that Gafarian et al [32] applied this method to an M/M/1/infinity queue, ρ = 0.1, 0.5, 0.7, 0.9 and found that it badly underestimated the truncation pt in almost all cases. Conway (1963) [40]: No testing carried out, just a brief description of method. Fishman (1973) [49]: No testing carried out, just a brief description of method. Bratley et al (1987) [50]: No testing carried out, but a brief description of method. They state that "Gafarian et al (1978) [32] show that if Conways rule (10) is applied to individual observations it will not work well."
Positives	Simple [32]. Recommended for use with low utilised systems [1]. Better than not truncating at all [5]. "For the empty queue experiments, Conway rule (10), gives reasonably accurate overall point estimates for

	the grand mean (within 3% of population mean). "The results for … [no truncation of data] and [Conway rule (10)] are identical to each other...and also provide the most precise interval estimate as expected. The results differ only on one run [out of 10] and provide the most accurate point estimate". For the loaded queue experiments, Conway rule (10) provide significant correction of bias through truncation. "For the filling queue experiments, ...the rule provides correction through truncation" [5].
Criticisms	"...poor approximation of the duration of the initial transient. As was shown in Gafarian et al. (1978) [32], using this rule we can significantly overestimate the length of the initial transient for small ρ and underestimate it for high ρ" [7]. Failed to find a truncation pt on many occasions (especially when run length was small) [32] . Reduces the bias much less than the Ensemble (Batch) Average Plots with Schribner's Rule (9) & Crossing-of-the-mean rule (12) [13]. "Gafarian et al (1978) [32] show that if Conway's rule is applied to individual observations it will not work well."[50] Requires preliminary (pilot) runs.
Accuracy	Poor
Simplicity	Good
Automation potential	Good
Generality (No assumptions)	Good
Parameters to estimate?	3 – number of replications to test as pilot runs, batch size, run length
Computation time	Good

Method ID	11
Method Name	Modified Conway Rule (or Backward Data-Interval Rule)
Brief description	Gafarian et al. (1978) [32]: "…turn Conway's idea around and continually look backwards to find the first observation that is neither a max nor min of all the previous observations. Thus number of observations in this procedure is a random variable, in contrast to the Conway rule (10)"
Literature Review	Preston White Jnr (1997) [5]: Systems Studied: Four M/M/1 queue systems with output variable set as number in system: 5. Empty Queue: model starts empty & idle with ρ = 0.9;

	6. Loaded Queue: 100 entities in system at start with ρ = 0.9; 7. Filling Tandem Queue: model starts empty & idle with ρ = 1.4 and a capacity of 100; 8. Transient Tandem Queue: model starts empty & idle with ρ = 1.4; Tested 4 methods: Conway rule (10), Modified Conway rule (11), Crossing-of-the-mean rule (12) and MCR (14). Results: Better than not truncating at all! For the empty queue experiments, un-weighted Modified Conway rule (11) gives inaccurate overall point estimates for the grand mean (outside 3% of population mean). For the loaded queue experiments, the rule provides significant correction of bias through truncation. For the filling queue experiments, the rule provides correction through truncation. Gafarian et al. (1978) [32] create this new method from the Conway rule (10). Systems Studied: M/M/1 queue with output variable as waiting time in queue per customer; ρ = 0.1, 0.5, 0.7, 0.9. Method is tested 100 times. 'Goodness' (Performance) Criteria: (vi) Accuracy: The ratio, mean of [estimated truncation values] / [true truncation value], is calculated and a value near 1 implies method accurate. (vii) Precision: Measure of variation in estimated truncation value is calculated as Sqrt(variance of estimated truncation points) / (mean of estimated truncation points). A value close to zero implies method precise. (viii) Generality: Judged to be general if the rule performs well across a broad range of systems and parameters within a system. (ix) Cost: Expense in computer time. (x) Simplicity: Accessible to average practitioner. (i) Accuaracy, (ii) Precision and (iii) Generality are considered first. Any method not satisfactory on all three is discarded. Computer cost is a last priority. Results: Modified Conway rule (11) satisfies simplicity criteria but fails the accuracy criteria, badly underestimating in almost all cases. Wilson&Pritsker (1978) [35]: Comments that Gafarian et al [32] applied this method to an M/M/1/infinity queue with ρ = 0.1, 0.5, 0.7, 0.9…and found it badly underestimated the truncation pt in almost all cases.

	Lee et al (1997) [58]: Systems Studied: M/M/1 and M/M/2 queuing systems with 4 levels of utilisation: 0.2, 0.5, 0.7 & 0.9. Performance criteria: Truncation pt; Coverage of true mean; Relative bias; Estimated relative half width of Confidence Interval. Results: "...in the case of ρ = 0.5 the Modified Conway rule (11) is best" when compared with the Crossing-of-the-mean rule (12) (with number of crossings set to 20), ED (45) and NN (46) methods.
Positives	Better than not truncating at all! [5] Simple [32]
Criticisms	Badly underestimates in almost all cases. [32] [35]
Accuracy	Poor
Simplicity	Good
Automation potential	Good
Generality (No assumptions)	Good
Parameters to estimate?	2 - run-length, number of pilot runs
Computation time	Good

Method ID	12
Method Name	Crossing-Of-The-Mean Rule
Brief description	Compute a running cumulative mean $\{c_1,c_2,\ldots,c_n\}$ as the data series $\{x_1,x_2,\ldots,x_n\}$ is generated. Calculate the total mean from all n observations. Then either i) count how many times the individual observations time series cross the total mean. (see refs [49][35][13] & [7]) ii) count how many times the individual observations cross the cumulative mean. (see ref [5]) iii) count how many times the cumulative mean crosses the total mean. (see ref [58]) (It is unclear in ref [32] which precise method they use.) In all cases choose truncation point as the point where the number of crossings reaches a specified number (k).

Literature Review	Mahajan & Ingalls (2004) [1]: Systems Studied: Job Shop model consisting of 5 Cells, C(i), i=1…5. Each cell has different number of machines (resources). There are 3 customer classes A,B,C. Overall arrival rate is poisson. Service times are exponential with mean dependent on customer class and cells. Arriving parts are split into classes {A, B, C} with probability {0.5, 0.3, 0.2}. 3 types of utilisation are used: TypeI is high utilisation with an average utilisation of 90% and range 80-95%; TypeII is moderate with an average utilisation of 70% and range 65-80%; TypeIII is Low with average utilisation of 50% with range 45-65%. Models are started empty & idle. Initial run length is 1000 hrs, with 5 replications. Performance criteria: Final MSE & Variance, average computing time, percentage change in MSE, percentage change in variance. Method said to perform well if it reduces both MSE and Variance and is computationally efficient. Results: Can be used with moderate to low utilised systems but requires a very long run length. Preston White Jnr (1997) [5]: Systems Studied: Four M/M/1 queue systems with output variable set as number in system: 9. Empty Queue: model starts empty & idle with ρ = 0.9; 10. Loaded Queue: 100 entities in system at start with ρ = 0.9; 11. Filling Tandem Queue: model starts empty & idle with ρ = 1.4 and a capacity of 100; 12. Transient Tandem Queue: model starts empty & idle with ρ = 1.4; Tested 4 methods: Conway rule (10), Modified Conway rule (11), Crossing-of-the-mean rule (12) and MCR (14). Results: Better than not truncating at all! For the empty queue experiments, Crossing-of-the-mean rule (12) gives reasonably accurate overall point estimates for the grand mean (within 3% of population mean). For the loaded queue experiments, this rule provides significant correction of bias through truncation. For the filling queue experiments, Crossing-of-the-mean rule (12) provides correction through truncation. Pawlikowski (1990) [7]: "This rule is sensitive to the value of k [set number of times crosses the mean]. Too large a value will usually lead to an overestimated value of [warm-up length] regardless of system's utilization, whereas too small a value can result in an underestimated [warm-up length] in more heavily loaded systems….. The system-dependent selection of the parameter k in this rule seems to be too arduous

	for potential users." Wilson & Pritsker (1978) [13]: Test Method: Used a 5 step procedure for evaluating 3 alternative start-up policies with 4 different truncation methods: IC1: initial conditions are idle & empty, IC2: start as close to steady state mode as possible, IC3: start as close to steady state mean as possible; Truncation Rules: TR0: retain all data, TR1: Use Conway rule (10), TR2: Use Crossing-of-the-mean rule (12), TR3: Use Ensemble (Batch) Average Plots with Schribner's Rule (9): Step1: Select standard system and compute bias, variance and MSE for all possible values of truncation point and for each initial condition (ICi) as specified above. Step2: Estimate distribution of truncation pt over independent simulation runs for each start up policy (ICi, TRi). Step3: Compute average bias, var & MSE by combining results from Steps1 & 2. Step4: Select a base policy & for every truncation pt under this policy compute the 1/2 length of a confidence interval (CI) for μ centred at the sample mean for that truncation pt. Average these values with respect to the observed distribution from Step2 to create a standard average CI 1/2 length for the base policy. Adjust CIs for all other policies so their average 1/2 length = the base average 1/2 length. Step5: Calculate probabilities that a CI centred around the estimated mean will actually cover the true mean. Calculate these coverage probabiities for all possible values of truncation pt, and average results w.r.t the appropriate truncation pt distribution for each policy. Systems Studied: M/M/1 queue with a capacity of 15 and a traffic intensity (ρ) of 0.9; A Machine Repair System - each machine in a group of 14 independent units has a constant failure rate of 0.2 so that time to failure is exponentially distributed with a mean of 5. Repair station rule is first come first served, with 3 parallel servers & enough waiting room for all machines. Repair times are exponentially distributed with a mean of 2. Arrival rate depends on how many units are in the system. Results: Crossing-of-the-mean rule (12) seemed to reduce the bias much more than the Conway rule (10) but suffered from a loss of CI coverage regarding the true mean. This showed apparently that a misspecification of the parameter/s can cause a large loss of coverage.

	Gafarian et al (1978) [32]: Systems Studied: M/M/1 queue with output variable as waiting time in queue per customer; ρ = 0.1, 0.5, 0.7, 0.9. Method is tested 100 times. 'Goodness' (Performance) Criteria: (xi) Accuracy: The ratio, mean of [estimated truncation values] / [true truncation value], is calculated and a value near 1 implies method accurate. (xii) Precision: Measure of variation in estimated truncation value is calculated as Sqrt(variance of estimated truncation points) / (mean of estimated truncation points). A value close to zero implies method precise. (xiii) Generality: Judged to be general if the rule performs well across a broad range of systems and parameters within a system. (xiv) Cost: Expense in computer time. (xv) Simplicity: Accessible to average practitioner. (i) Accuaracy, (ii) Precision and (iii) Generality are considered first. Any method not satisfactory on all three is discarded. Computer cost is a last priority. Results: Crossing-of-the-mean rule (12) satisfied the simplicity criteria and wastes no data. Regarding the accuracy criteria, it is very conservative for low ρ, i.e. overestimates truncation pt. Precision improves with decreasing ρ and increasing number of crossings. Wilson & Pritsker (1978) [35]: Some guidelines for parameter selection found in Wilson 1977. Gafarian et al [32] applied this method to an M/M/1/infinity queue with ρ = 0.1, 0.5, 0.7, 0.9 and found it badly overestimated the truncation pt for low values of ρ. It was better for higher values of ρ but the precision of the estimated mean decreased. Lee et al (1997) [58]: Systems Studied: M/M/1 and M/M/2 queuing systems with 4 levels of utilisation: 0.2, 0.5, 0.7 and 0.9. Performance criteria: Truncation pt; Coverage of true mean; Relative bias; Estimated relative half width of CI. Results: "...in case of ρ = 0.2 and n(the number of batches) = 50, the Crossing-of-the-mean rule (12) (with crossings set to 20) is superior to the Modified Conway rule (11), ED (45) and NN (46) methods." For ρ =

	0.2 and n = 1000, Crossing-of-the-mean rule (12) or ED (45) method is superior. Fishman (1973) [49]: Explains rule as an initialisation bias test rather than a straight truncation method.
Positives	Simple and wastes no data [32]. Better than not truncating at all![5] Reduces the bias much more than Conway rule (10) [13]. Precision improves with decreasing ρ and increasing number of crossings [32]. Performs better for higher values of ρ but precision of estimated mean decreased [35]. Can be used with moderate to low utilised systems…[1]
Criticisms	…but requires a very long run length. [1] Rule is sensitive to the value of k (set number of times crosses the mean) which can lead to over or under estimation of warm-up length.[7] "The system-dependent selection of the parameter k in this rule seems to be too arduous for potential users." [7] Suffered from a loss of CI coverage regarding the true mean…apparently a misspecification of the parameter/s can cause a large loss of coverage. [13] Badly overestimated truncation pt for low values of ρ. [32] [35].
Accuracy	Poor
Simplicity	Good
Automation potential	Good
Generality (No assumptions)	Good
Parameters to estimate?	2 – run length (n) and number of crossings (k)
Computation time	Good

Method ID	13
Method Name	Autocorrelation Estimator Rule
Brief description	Pawlikowski (1990) [7]: "Fishman proposed equating the variance of the mean of auto-correlated observations with the variance of the mean of a hypothetical sequence of independent observations to find the number of collected (auto-correlated) observations equivalent, in the above sense, to one independent (hypothetical) observation. After some simplification we get the following rule: In a time series of observations $x_1, x_2, \ldots, x_n, \ldots$, the initial transient is over after $n_0 = 2\sum_{k=1}^{n-1}\left(1-\frac{k}{n}\right)\frac{\hat{R}(k)}{\hat{R}(0)}$ observations, where $\hat{R}(k)$ is the estimator of the autocorrelation of the lag k,

	0≤k≤n-1.”
Literature Review	Pawlikowski (1990) [7]: "…usually gives underestimates [of warm-up length]" Fishman (1971) [24]: Systems Studied: Zero, 1st, and 2nd order Autoregressive schemes for normal stochastic sequences. M/M/1 queue, with ρ = 0.9. Output variable is mean queue length. Performance criteria: Number of times generated CIs included true mean, which is 9. Results: Poor with smaller n. Performance increases with increasing data (n). Poor results for M/M/1 model as tested. Wilson& Pritsker (1978) [35]: Brief explanation of method - no comments made.
Positives	Performance increases with increasing data (n). [24]
Criticisms	Usually gives underestimates of warm-up length [7]. Poor with smaller n. [24]
Accuracy	Poor
Simplicity	Medium
Automation potential	Good
Generality (No assumptions)	Good
Parameters to estimate?	1 - Autocorrelation of lag k (V)
Computation time	?

Method ID	14
Method Name	Marginal Confidence Rule (MCR) or Marginal Standard Error Rule (MSER)
Brief description	Preston White Jnr (1997) [5]: “Instead of selecting a truncation point to minimise the MSE, we propose to select a truncation point that minimises the width of the CI about the truncated sample mean …Thus we will seek to mitigate bias by removing initial observations that are far from the sample mean, but only to the extent this distance is sufficient to compensate for the resulting reduction in sample size in the calculation of the confidence interval half-width.” From refs [36] & [5]: Formally, given a finite stochastic sequence of output i of replication j $\{Y_i(j)$:

	i=1,2,…,n}, we define the optimal truncation point for this sequence as: $d(j)^* = \underset{n>d(j)\geq 0}{\arg\min}\left[\frac{z_{\alpha/2}\, s(d(j))}{\sqrt{n(j)-d(j)}}\right]$, where $z_{\alpha/2}$ is the value of the N(0,1) distribution associated with a 100(1-α)% CI and s(d(j)) is the sample standard deviation of the reserved sequence (i.e. of all data following d(j)), and n(j) is the total number of observations in replication j. Since the confidence level α is fixed, $z_{\alpha/2}$ is a constant and can therefore be set arbitrarily to 1, as the purpose of using the above equation is only to compare all data points to find the minimum. The expression for the optimal truncation point can therefore be written explicitly in terms of the observations: $d(j)^* = \underset{n>d(j)\geq 0}{\arg\min}\left[\frac{1}{(n(j)-d(j))^2}\sum_{i=d+1}^{n}\left(Y_i(j)-\bar{Y}_{n,d}(j)\right)^2\right]$
Literature Review	Preston White Jnr (1997) [5]: Systems Studied: Four M/M/1 queue systems with output variable set as number in system: 13. Empty Queue: model starts empty & idle with ρ = 0.9; 14. Loaded Queue: 100 entities in system at start with ρ = 0.9; 15. Filling Tandem Queue: model starts empty & idle with ρ = 1.4 and a capacity of 100; 16. Transient Tandem Queue: model starts empty & idle with ρ = 1.4; Tested 4 methods: Conway rule (10), Modified Conway rule (11), Crossing-of-the-mean rule (12) and MCR (14). Results: All better than not truncating at all! MCR (14) was comparable or slightly superior to other tests used here. MCR is easy to understand & implement, inexpensive to compute, efficient in preserving data & efficient in mitigating initial bias. "For the empty queue experiments, all of the rules give reasonably accurate overall point estimates for the grand mean (within 3% of population mean), with the sole exception of the un-weighted [Modified Conway rule (11)]. Only MCR (14) provides coverage of the 95% confidence interval however." "For the loaded queue experiments, ...all of the rules provide significant correction [of bias] through truncation. Again only the MCR yields coverage of the 95% CI." "For the filling queue experiments, ...the rules provide correction through truncation, with remarkably consistent point and interval

	estimates of the grand sample mean across rules and across runs." "…while removing rare but recurring observations improves point estimates of the mean, truncation of unusual observations aggravates underestimation of the steady-state variance of the output, if very few runs are made. This is inevitable and methods such as batch means will continue to be required in developing interval estimates of the mean for such applications". Preston White Jnr et al (2000) [28]: Systems Studied: 2nd order autoregressive process with zero mean and with differing initial parameter values. 3 bias fns: exponential, mean shift, under-damped oscillations (based on Cash et al (1992) [33]). Bias fns incorporated by superposition (adding into output) and injection (adding into state equation). Performance Criteria: Sample mean, abs bias = |grand estimated mean - grand mean of unbiased data|; p-value of 2 sample t-test - H0: estimated mean = mean from unbiased data; average computation time; min, max, mean & standard deviation of truncation pt.; number of inconclusive results. The MCR/MSER method is applied sequentially. Results: MCR (14) consistently outperformed the other methods: Schruben's modified test (21), BM Max test (24) and BM Batch means test (25), except for the MSER-5 (15) method on models with exp and mean shift bias. Highly accurate in locating optimal truncation pt. Not as effective with damped oscillating bias. Performance decreased with increasing average bias (possibly due to sensitivity to individual observations). Effective at detecting mean shift bias in capacitated data sets. MSER-5 (15) performed best. The methods MCR (14), Schruben's modified test (21) and BM Max test (24) are much of a much-ness. BM Batch means test (25) performed the worst. Linton & Harmonsky (2002) [36]: Model1: Queuing model with 2 servers in sequence. Inter-arrival time is exponential with mean of 8; server1 process time is exponential with a mean of 6; server2 process time is exponential with mean of 7. Model2: Same as for model1 except all distributions are triangular; inter-arrival time distribution is therefore defined as{min = 6, mode = 8, max = 10}; server1 process time defined as{min = 4, mode = 6, max = 8}; server2 process time as {min = 5, mode = 7, max = 9} Performance criteria: i) equality of variance (between methods) using Levene's test

	ii) equality of mean (between methods) using 2-sample t-test. Results: MCR (14) is intuitively appealing because it determines a truncations point for each replication. It was found to be able to adjust to changes in distributions of the inter-arrival times and processing times. Computationally intensive as requires a large number of calculations. WARNING: This paper did not test if chosen truncation point (& therefore estimate of mean output) was correct or efficient!
Positives	MCR (14) is intuitively appealing because it determines a truncations point for each replication. Able to adjust to changes in distributions of the inter-arrival times and processing times [36]. Highly accurate in locating optimal truncation pt for models with exp and mean shift bias. Effective at detecting mean shift bias in capacitated data sets [28]. Performance of MCR was either superior or on par with methods Conway rule (10), Modified Conway rule (11), Crossing-of-the-mean rule (12), Schruben's modified test (21), BM Max test (24) and BM Batch means test (25). (from refs[5] & [28])
Criticisms	Not as effective with damped oscillating bias. Performance decreased with increasing average bias (possibly due to sensitivity to individual observations) [28]. Computationally intensive as requires a large number of calculations. [36]
Accuracy	Medium/Good
Simplicity	Good
Automation potential	Good
Generality (No assumptions)	Good
Parameters to estimate?	1 – run length (n)
Computation time	Medium

Method ID	15
Method Name	Marginal Standard Error Rule m (MSER–m) (e.g. m=5)
Brief description	Same as for MCR/MSER(14) but MSER-m applies the equation $d(j)^* = \underset{n>d(j)\geq 0}{\arg\min}\left[\frac{1}{(n(j)-d(j))^2}\sum_{i=d+1}^{n}\left(Y_i(j)-\bar{Y}_{n,d}(j)\right)^2\right]$, (see method (14)), to a series of $b=\left\lfloor n/m \right\rfloor$ batch averages instead of to the raw output series.

Literature Review	Mahajan & Ingalls (2004) [1]: Systems Studied: Job Shop model consisting of 5 Cells, C(i), i=1…5. Each cell has different number of machines (resources). There are 3 customer classes A,B,C. Overall arrival rate is poisson. Service times are exponential with mean dependent on customer class and cells. Arriving parts are split into classes {A, B, C} with probability {0.5, 0.3, 0.2}. 3 types of utilisation are used: TypeI is high utilisation with an average utilisation of 90% and range 80-95%; TypeII is moderate with an average utilisation of 70% and range 65-80%; TypeIII is Low with average utilisation of 50% with range 45-65%. Models are started empty & idle. Initial run length is 1000 hrs, with 5 replications. Performance criteria: Final MSE & Variance, average computing time, percentage change in MSE, percentage change in variance. Method said to perform well if it reduces both MSE and Variance and is computationally efficient. Results: Recommended MSER-5 for use with low to highly utilised systems with a long run length. Preston White Jnr et al (2000) [28]: Systems Studied: 2nd order autoregressive process with zero mean and with differing initial parameter values. 3 bias fns: exponential, mean shift, under-damped oscillations (based on Cash et al (1992) [33]). Bias fns incorporated by superposition (adding into output) and injection (adding into state equation). Performance Criteria: Sample mean, abs bias = |grand estimated mean - grand mean of unbiased data|; p-value of 2 sample t-test - H0: estimated mean = mean from unbiased data; average computation time; min, max, mean & standard deviation of truncation pt.; number of inconclusive results. The MCR/MSER method is applied sequentially. Results: MSER-5 (15) most effective and robust method when compared with methods: Schruben's modified test (21), BM Max test (24), BM Batch means test (25) & MCR (14). Particularly effective with exponential and mean shift bias. Unlike MCR (14), if bias increased so did MSER-5's effectiveness. Also fastest method especially on data sets with big bias. MSER-5 (15) performed best. The methods MCR (14), Schruben's modified test (21) and BM Max test (24) are much of a much-ness. BM Batch means test (25) performed the worst. Sandikci & Sabuncuoglu (2006) [45]: Systems Studied: 2 types of manufacturing system: (1) serial production lines; (2) job-shops. Output

	variable is time in system. Results: "MSER-5 applied to the whole sequence suggests truncating 4876 observations whereas deleting [outliers] from the sequence would change the truncation point drastically to 339. This shows that unless extreme values are carefully deleted from a sequence MSER can display a poor performance." "Since the MSER is an objective criterion…and is very simple and computationally efficient, we recommend this heuristic. However special attention must be paid to remove any outliers from the sequence which otherwise would lead the analysts to wrong conclusions" MSER-5 (15) had "comparable" results with Cumulative Mean Rule (3). "Cumulative averages usually suggest longer transient periods than MSER-5"
Positives	Recommended MSER-5 for use with low to highly utilised systems with a long run length [1]. Effective and robust. If bias increased so did MSER-5's effectiveness. Particularly effective with exp and mean shift bias. Fast method especially on data sets with big bias [28]. Objective criterion…and is very simple and computationally efficient [45].
Criticisms	"…unless extreme values are carefully deleted from a sequence MSER can display a poor performance."[45]
Accuracy	Good
Simplicity	Good
Automation potential	Good
Generality (No assumptions)	Good
Parameters to estimate?	2 – run length, batch size (although this could be set to 5 by default)
Computation time	Medium

Method ID	17
Method Name	Relaxation Heuristics
Brief description	Linton & Harmonsky (2002) [36] & Roth (1994) [6]: For M/M/k systems: Begin queuing system at rest and empty. Truncation point = $4\tau_R$ units of model time, where: $\tau_R = \left[1.4k\mu(1-\rho)^2\right]^{-1}$, where k = number of servers, μ = mean service rate and ρ = traffic intensity. $k \le k_{max}$ *(max number of servers which is dependent on ρ: see Roth(1994)* [6]*)*

	Roth & Josephy(1993) [57]: For $M/E_k/1$ & $E_k/M/1$ systems, begin queuing system at rest and empty. Truncation point = $4\tau_R$ units of model time, where: $\tau_R = \frac{1+ {}^1\!/\!_k}{2.8\mu(1-\rho)^2}$, where k = number of servers, μ = service rate and ρ = traffic intensity.
Literature Review	Roth (1994) [6]: Systems Studied: 51 M/M/k queuing systems with ρ = 0.1 to 0.9 and k = 1 to max. Performance Criteria: Normal hypothesis test – H0: estimated mean = true mean (see paper [6] for details). Results: Relaxation Heuristics (17) satisfies performance criteria in all except M/M/2 ρ = 0.2; Cumulative-mean rule (3) fails twice for ρ = 0.9; Kelton-Law regression method (18) & Truncated mean rule (4) failed in 43% & 29% of cases respectively - probably because run length was not long enough.. Pawlikowski (1990) [7]: "For more complex queuing networks the relaxation times have not yet been theoretically determined." "…the usefulness of even known formulas for relaxation times can be questioned in simulation studies. They can be used only as first approximations of the duration of simulated initial transients, since it has been shown that estimators of the mean values from simulation tend to their steady state more slowly than exponentially.…" Linton & Harmonsky (2002) [36]: Model1: Queuing model with 2 servers in sequence. Inter-arrival time is exponential with mean of 8; server1 process time is exponential with a mean of 6; server2 process time is exponential with mean of 7. Model2: Same as for model1 except all distributions are triangular; inter-arrival time distribution is therefore defined as{min = 6, mode = 8, max = 10}; server1 process time defined as{min = 4, mode = 6, max = 8}; server2 process time as {min = 5, mode = 7, max = 9} Performance criteria: i) equality of variance (between methods) using Levene's test ii) equality of mean (between methods) using 2-sample t-test. Results: Produced comparable results with Welch's method (6). The Relaxation Heuristics (17) were found able to adjust to changes in distributions of the inter-arrival times and processing times. Does not seem to be

	negatively affected by modifications necessary to apply it to the chosen two models. A more practical method. WARNING: Not tested in this paper to see if chosen truncation point was correct or efficient! Kimbler & Knight (1987) [46]: Systems Studied: M/M/1 queue with ρ = 0.9. This is a "highly congested system …has an unusually gradual transition into steady state" Performance Criteria: Steady state statistics (time in system) compared with theoretical solution using Dudewicz method for determining the best of k systems. 4 Truncation methods tested: Kelton-Law regression method (18), Optimal test (22), Kimbler's double exponential smoothing (43) and Relaxation Heuristics (17). Results: Relaxation Heuristics (17) had the largest average truncation pt of the four methods tested and was only 8% below the theoretical mean after 20 replications. Very simple method but only applicable (as presented here) to M/M/1 system for which an analytical solution exists. It is the simplest method of the four tested but with limited applicability. Roth & Josephy (1993) [57]: Extends the scope of the Relaxation Heuristics (17) to include M/Ek/1 and Ek/M/1 queueing systems. Systems Studied: 30 M/Ek/1 and 24 Ek/M/1 queuing systems (which begin at rest); ρ varies from 0.25 to 0.925 by holding arrival rate fixed and varying service rate; k (parameter for the Erlang dist) varies from 1 to 20. Note: Experiments have fixed replication length. Performance Criteria: Bias = estimated mean - theoretical mean; CI size - a function of the variance; MSE; observed coverage probability of CIs. Results: "The relaxation time heuristic satisfies the bias criterion in each experiment. The Kelton-Law regression method (18), Cumulative-mean rule (3) and Truncated mean rule (4) techniques are less consistent, causing rejection of the null hypothesis in 33, 10, and 3% of cases respectively." "The confidence interval coverage probabilities are quite consistent for each truncation rule." "Of …concern is performance of the relaxation time heuristic when Erlang order k (in Ek/M/1) is large and traffic intensity ρ is low. Recall that the relaxation time (TR) is an approximation. It is clear that the Odoni and Roth approx for TR is not as appropriate for these systems as it is for the other cases [tested]. In practice one would use a larger multiplier of TR [i.e. larger than 4] for such systems. Future refinement of the approx should alleviate

	this prob." "The major advantage of the relaxation heuristic is that it only requires specification of model characteristics; there are no input parameters left for the user to specify. This eliminates the need for preliminary replications and as there is no randomness in the truncation pt selection, the user may have confidence that bias is removed in a consistent manner."
Positives	Produced comparable results with Welch's method (6). Able to adjust to changes in distributions of the inter-arrival times and processing times. Does not seem to be negatively affected by modifications necessary to apply it to the chosen two models. A practical method.[36] A relatively simple method [46]. "Only requires specification of model characteristics; there are no input parameters to estimate". "..there is no randomness in the truncation pt selection, the user may have confidence that bias is removed in a consistent manner."[57]
Criticisms	"…the usefulness of even known formulas for relaxation times can be questioned in simulation studies. They can be used only as first approximations of the duration of simulated initial transients, since it has been shown that estimators of the mean values from simulation tend to their steady state more slowly than exponentially…". "For more complex queueing networks the relaxation times have not yet been theoretically determined." [7] Has limited applicability.[46] Of concern is performance of the relaxation time heuristic when Erlang order k (in Ek/M/1) is large and traffic intensity ρ is low….The relaxation time (TR) is an approximation and is not as appropriate for these systems as it is for the other cases [tested]. In practice one would use a larger multiplier of TR [i.e. larger than 4] for such systems. Future refinement of the approx should alleviate this prob." [57]
Accuracy	Good/Medium
Simplicity	Good
Automation potential	Poor/Medium – requires input values from system and being able to recognise the type of system so that the correct relaxation time equation can be applied if one exists.
Generality (No assumptions)	Poor
Parameters to estimate?	None
Computation time	Good

Method ID	29
Method Name	Exponentially Weighted Moving Average Control Charts
Brief description	Rossetti et al (2005)[9]: Let L =Average run length constant; λ = smoothing constant; d = deletion point and p_2' = desired proportion of the exp weighted data remaining after deletion that fall within control limits. The exponentially weighted moving average is defined as: $Z_i = \lambda Y_i + (1-\lambda)Z_{i-1}$, where $0<\lambda<1$ and the starting value $Z_0=\mu_0$. In an EMWA control chart the control limits are calculated by: $UCL = \mu_0 + L\sigma\sqrt{\frac{\lambda}{(2-\lambda)}\left[1-(1-\lambda)^{2i}\right]}$ $CenterLine = \mu_0$ $LCL = \mu_0 - L\sigma\sqrt{\frac{\lambda}{(2-\lambda)}\left[1-(1-\lambda)^{2i}\right]}$ Where σ is the variance of the underlying data. Define an indicator variable to indicate whether an exp weighted point is in control or not: $I_i(d) = \begin{cases} 1 & \text{if } Z_i \in \left[LCL(d), UCL(d)\right] \\ 0 & \text{otherwise} \end{cases}$ Therefore an estimator for p_2' can be defined as : $\hat{p}_2(d) = \frac{\sum_{i=d+1}^{n} I_i(d)}{n-d}$ The procedure is : 1. Collect the data $\{Y_i;\ i=1,2,\ldots,n\}$ and reverse the data.

	2. Set p_2', λ and L 3. Set a = 0 and b = n 4. Set d = Int[(a+b)/2] 5. If $\hat{p}_2(d) < p_2'$ then a = a, b = d Else a = d, b = b 6. If a = b return d_r = a Else goto step 4.
Literature Review	Rossetti et al (2005)[9]: Systems Studied: Artificial data sets using a monotonically decreasing function & a monotonically exponential decreasing function for bias. M/M/1 queue with ρ = 02 & 0.8. Output variable is waiting times; Performance measures: bias, variance & MSE of point estimator. Results: As the percentage of biased data points increases this method consistently underestimates the warm-up period. Also performs worse when variance increases. Autocorrelation is detrimental to the procedure but can be helped by batching. "It is easy to implement and relatively quick computationally".
Positives	Easy to implement. Relatively quick computationally [9].
Criticisms	As the percentage of biased data pts increases the method consistently underestimates the warm-up period. More prone to underestimation as variance (white noise of the process) increases. Autocorrelation detrimental to procedure (but can be helped by batching). [9]
Accuracy	Poor (prone to underestimate)
Simplicity	Medium
Automation potential	Medium/Good
Generality (No assumptions)	Good
Parameters to estimate?	5 - Run length (n); Average run length constant (L); smoothing constant (λ); desired proportion of the exp weighted data remaining after deletion that fall within control limits (p_2'); Variance (V)
Computation time	Good

Method ID	40
Method Name	Chaos Theory Methods (M1 and M2)
Brief description	Lee & Oh (1994) [42]: M1: Let $\lambda_i = \frac{1}{k}\sum_{i=1}^{k}\log_2 \frac{x_{i+1}}{x_i}$, k = 1,2,…,n-1 where x_i = average waiting time in queue for up to the i[th] customer from the 1[st] customer in M/M/1 with arrival rate λ, service rate μ. n = run length The criteria for λ_i to decide the truncation point are: C1: $\|\lambda_i\|$ must be less than the specified value The specified values for variation rate 5% is obtained by: $\log_2 \frac{[a(1 \pm 0.025)]}{a} = 0.036$ C2: λ_i must not have the positive values (or negative values) more than 30 times continuously. M2: Partition n simulation data $x_1, x_2, \ldots, x_n$ into b nonoverlapping batches in which each batch has m observations such that n=b*m, and define the following functions of the orig data for I = 1,2,…,b. $\lambda_i = \frac{1}{m}\sum_{j=1}^{m}\log_2 \frac{x?}{x?}$, can not read from paper (or screen!) Where x_i = average waiting time in queue for up to the j[th] customer in the i[th] batch. C3: $\|\lambda_i\|$ must be less than the specified value

	The specified values for variation rate 5% is obtained by: $\log_2 \frac{[a(1 \pm 0.025)]}{a} = 0.036$
Literature Review	Lee & Oh (1994) [42]: Systems Studied: M/M/1/infinity queuing model with ρ = 0.5; 0.7; 0.9. Output variable is waiting time in queue. Performance criteria: Effectiveness was assessed using a bias calculation: Abs[true mean - estimated mean] / estimated mean Results: M2 (40) with a batch size of 40 worked best. For ρ = 0.5 and 0.7, methods BM max test (24), M1 & M2 (40) had results similar. For ρ = 0.9, with given present run-length, system fails to reach steady state in some runs. Methods M1 (40) and BM max test (24) "detect a truncation pt regardless of system state". But M2 (40) "can determine correctly the transient state…" "Some statistical work is required to determine the most appropriate criteria to be used in the proposed methods M1 & M2 (40)." M2 (40) of the new methods seemed to perform better than the BM max test (24).
Positives	M2 (40) "can determine correctly the transient state…" [42]
Criticisms	"Some statistical work is required to determine the most appropriate criteria to be used in the proposed methods." [42]
Accuracy	M1: Poor; M2: Medium
Simplicity	Good
Automation potential	Good
Generality (No assumptions)	Poor - applied to estimating average waiting time in M/M/1 queue model
Parameters to estimate?	2 - Run length (n); batch number (b)
Computation time	Good

Method ID	41
Method Name	Beck's Approach for Cyclic Output
Brief description	"Involves comparing the values for each hour in the 1st week of the simulation run, against the values in the subsequent weeks for the same hour."

	“The following criterion was used to specify when a model had warmed-up: For 3 consecutive hours the values had to be non-zero and also not greater than the maximum values or less than the minimum values in the subsequent weeks.
Literature Review	Beck (2004) [44]: Uses a British Airways model called "Operations Robustness Model", flies a schedule for a number of weeks and outputs a matrix of operational performance figures. Produces a cyclic output to illustrate method.
Positives	-
Criticisms	-
Accuracy	Unknown
Simplicity	Good
Automation potential	Medium
Generality (No assumptions)	Poor - Cyclic data (with a regular cycle?)
Parameters to estimate?	None
Computation time	Good

Method ID	42
Method Name	Tocher’s Cycle Rule
Brief description	“The performance of a system can be regarded as a cyclic evolution of system’s basic operations. For this reason, Tocher (1963, p.176) suggested this rule: The initial transient period is over if the longest cycle distinguished in the behaviour of the simulated system has been executed at least 3 or 4 times.”
Literature Review	Pawlikowski (1990) [7]: "No results concerning the effectiveness of this rule are available."
Positives	-
Criticisms	How to find a system cycle? And monitor it?
Accuracy	unknown
Simplicity	Good
Automation potential	Poor
Generality (No assumptions)	Good

Parameters to estimate?	None
Computation time	Good

Method ID	43
Method Name	Kimbler's Double Exponential Smoothing
Brief description	Kimbler & Knight (1987) [46]: Calculate the doubly-smoothed sequence of waiting times using: $F(t) = \begin{cases} \alpha W(t) + (1-\alpha)F(t-1), & t>1 \\ W(t), & t=1 \end{cases}$ $\hat{W}(t) = \begin{cases} \alpha F(t) + (1-\alpha)\hat{W}(t-1), & t>1 \\ W(t) & t=1 \end{cases}$ Where *F(t)* is the smoothed sequence of waiting times, and $\hat{W}(t)$ is the doubly-smoothed sequence of waiting times. Calculate the error *e(t)*, smoothed error *E(t)* and absolute smoothed error *A(t)* using: $E(t) = \begin{cases} \lambda e(t) + (1-\lambda)E(t-1), & t>1 \\ 0, & t=1 \end{cases}$ $A(t) = \begin{cases} \lambda \lvert e(t) \rvert + (1-\lambda)A(t-1), & t>1 \\ 0 & t=1 \end{cases}$ Where $e(t) = W(t) - \hat{W}(t)$ The tracking signal is: TS = E(t) / A(t), t>1 Perform these calculations for each subsequent time series value and compare to a constant C. As long as the tracking signal is larger in absolute value than C, the process is assumed to be in the transient period. Suggested parameter values: α = 0.1; λ = 0.6; C = 0.1

Literature Review	Kimbler & Knight (1987) [46]: Systems Studied: M/M/1 queue with ρ = 0.9. This is a "highly congested system …has an unusually gradual transition into steady state" Performance Criteria: Steady state statistics (time in system) compared with theoretical solution using Dudewicz method for determining the best of k systems. Tests 4 methods: Relaxation heuristics (17), Optimal test (22), Kelton & Law regression (18), Kimbler's Double Exponential Smoothing (43). Results: The Double exponential smoothing method had the smallest average truncation pt of the four methods tested. It is simple to execute & worked well. "We feel the best of the four methods was the one developed by Kimbler. It worked as well at bias removal as any of the other three and was very easy to apply. An additional advantage which it had over the methods of Kelton & Law regression (18) and Schruben (Optimal test (22)) was its ability to work from the beginning to the end of the time sequence. Unlike the other methods which start at the end of the time sequence work back toward the beginning, the Kimbler method is able to determine truncation points 'on the fly' without the need for a pilot run." All methods worked equally well in relation to estimating mean time in system...but there were quite a difference in truncation points. Kimbler's Double Exponential Smoothing (43) had the smallest average truncation pt of the four methods tested. Relaxation heuristics (17) and Kimbler's Double Exponential Smoothing (43) are far simpler than Kelton & Law regression (18) and Optimal test (22) to carry out.
Positives	Simple to carry out. "…able to determine truncation points 'on the fly' without the need for a pilot run." Had the smallest average truncation pt of the four methods tested, Relaxation heuristics (17), Optimal test (22), Kelton & Law regression (18), Kimbler's Double Exponential Smoothing (43). "It worked as well at bias removal as any of the other three." "…worked equally well in relation to estimating mean time in system." [46]
Criticisms	-
Accuracy	Good/Medium
Simplicity	Good
Automation potential	Good

Generality (No assumptions)	? – unclear from paper whether this method should only be applied to waiting times
Parameters to estimate?	$\alpha = 0.1$; $\lambda = 0.6$; $C = 0.1$
Computation time	Good

Method ID	45
Method Name	Euclidean Distance Method (ED)
Brief description	Lee et al (1997) [58]: Collect 10 data points; Normalise this vector of 10 data pts; check ED criteria; if criteria satisfied 5 times in a row (i.e. for 5 lots of 10 data points in a row) then truncation pt determined. Normalise by: dividing vector of 10 data pts ($x_1, x_2,\ldots,x_{10}$) by $\sqrt{x_1^2 + x_2^2 + \ldots + x_{10}^2}$ ED Criteria: Mean of the normalised data lies between 0.3162276 and 0.3162278.
Literature Review	Lee et al (1997) [58]: Systems Studied: M/M/1 and M/M/2 queuing systems with 4 levels of utilisation: 0.2, 0.5, 0.7 & 0.9. Performance criteria: Truncation pt; Coverage of true mean; Relative bias; Estimated relative half width of Confidence Interval. Four methods tested: Crossing-Mean-Rule (12), Modified Conway rule (11), ED (45) & NN (46) methods. Results: In the case of $\rho = 0.2$ and n (number of batches) = 50, Crossing-Mean-Rule (12) with crossings set to 20, is superior to others tested. For $\rho = 0.2$ and n = 1000, Crossing-Mean-Rule (12) or ED (45) method is better. For $\rho = 0.5$, Modified Conway rule (11) is best. For $\rho = 0.7$, NN (46) Method is better than any of the others. For $\rho = 0.9$, ED (45) & NN (46) methods are better w.r.t coverage and relative bias. "Regardless of ρ ED (45) & NN (46) have robustness in their performance." Batching was used to find the estimated mean after truncation was achieved - "batch size and number of batches …must be determined on-line for completely automated simulation output analysis. One of the methods for this purpose would be sequential procedure suggested by Law and Kelton [52]. If we use sequential procedure after determining truncation pt by ED or NN, we could achieve this goal".
Positives	For $\rho = 0.9$, ED (45) & NN (46) methods are better w.r.t coverage and relative bias…Regardless of ρ, ED&

	NN have robustness in their performance." [58]
Criticisms	Do not perform so well with low traffic intensity.
Accuracy	Medium
Simplicity	Good
Automation potential	Good
Generality (No assumptions)	Good
Parameters to estimate?	None
Computation time	Good

Method ID	46
Method Name	Neural Networks (NN) method
Brief description	Lee et al (1997) [58]: Training method: Input to Neural Network are10 normalised data points and 0 (or 1) for identifying steady-state of normalised data. Data from an M/M/1 simulation with $\rho = 0.2$ is used for training. Using the weights determined by supervised learning, the NN receives 10 data points as input and gives 0 or 1 as output. 1 means that the steady-state pattern of training data coincides with the current pattern within a limited error range. NN criterion is met when output is 1. Consider warm-up period over when NN criterion is satisfied 5 times successively.
Literature Review	Lee et al (1997) [58]: Systems Studied: M/M/1 and M/M/2 queuing systems with 4 levels of utilisation: 0.2, 0.5, 0.7 & 0.9. Performance criteria: Truncation pt; Coverage of true mean; Relative bias; Estimated relative half width of Confidence Interval. Four methods tested: Crossing-Mean-Rule (12), Modified Conway rule (11), ED (45) & NN (46) methods. Results: In the case of $\rho = 0.2$ and n (number of batches) = 50, Crossing-Mean-Rule (12) with crossings set to 20, is superior to others tested. For $\rho = 0.2$ and n = 1000, Crossing-Mean-Rule (12) or ED (45) method is better. For $\rho = 0.5$, Modified Conway rule (11) is best. For $\rho = 0.7$, NN (46) Method is better than any of the others. For $\rho = 0.9$, ED (45) & NN (46) methods are better w.r.t coverage and relative bias. "Regardless of ρ ED (45) & NN (46) have robustness in their performance." Batching was used to find the estimated

	mean after truncation was achieved - "batch size and number of batches …must be determined on-line for completely automated simulation output analysis. One of the methods for this purpose would be sequential procedure suggested by Law and Kelton [52]. If we use sequential procedure after determining truncation pt by ED or NN, we could achieve this goal".
Positives	For $\rho = 0.9$, ED (45) & NN (46) methods are better w.r.t coverage and relative bias…Regardless of ρ, ED& NN have robustness in their performance." [58]
Criticisms	-
Accuracy	Medium
Simplicity	Poor
Automation potential	Poor
Generality (No assumptions)	Good
Parameters to estimate?	None
Computation time	unknown

(Return to top of document)

INITIALISATION BIAS TESTS

Method ID	20
Method Name	Schruben’s Maximum Test (STS)
Brief description	Schruben (1982) [20]: 1. Let $Y_1, Y_2, \ldots, Y_m$ be a series of output from a run of length m. 2. Let $S_m(k)$ = total mean of series. 3. Let s^* be the 1st (if there are ties) global maximum of $(k\,S_m(k)/\sqrt{m})$, for $k = 1,2,\ldots,m$. Let k^* be the location of this maximum. 4. Let $t^* = k^*/m$ 5. Let τ^{*2} be the estimate of the variance $V = [\bar{Y}(m)]$. 6. Calculate the test statistic: $h^* = s^{*2}/[3\,\tau^{*2} t^*(1-t^*)]$, which has an approx F dist with 3 and δ degrees of

	freedom (see ref[20] for more about estimating the value of δ). 7. Test the null hypothesis of no negative bias with this F test. If wanting to test for positive bias should use opposite sign for test statistic. This method can be carried out for one run or averaged replications.
Literature Review	Schruben et al(1983) [18]: Systems Studied: Model1: AR(2) with oscillating initial transient function. Model2: M/M/1 queue with $\rho = 0.9$. Output variable is waiting times. Model3: Simple production-inventory model. Model4: time-shared computer system (Adiri & Avi-Itzhak [1969]). Model5: Telephone exchange model (M/M/s). Output is number of busy lines per time period. Model6: Network of 3 capacitated M/M/s queues with feedback. Performance criteria: Percentage of runs for which the null hypothesis of no bias was rejected. Results: Optimal test (22) has high power in detecting presence of initialisation bias in all test models. Compared well with the Schruben Max test (20), which performed poorly for 2nd model (M/M/1 $\rho = 0.9$). "The computation involved in the two tests was minimal even on the smallest of computers". "We recommend that both the test presented here, Optimal test (22), and the test presented by Schruben, Schruben Max test (20), be conducted to support ascertations that initialisation bias in the mean of a simulation output process has been effectively controlled." Schruben (1982) [20]: Systems Studied: Model1: Telephone Exchange (McDaniel [1979]). Model2: Inventory system. Model3: Computer time sharing system (Adiri & Avi-Itzhak [1969], Sargeant [1979]). Model4: M/M/1 queue, with $\rho = 0.9$. Model5: Network of 3 capacitated M/M/s queues with feedback. Performance criteria: For the test procedure to be valid the significance level (alpha) should be uniformly (0, 1) distributed when no initial bias is present.

	Effectiveness: The power of the test has the distribution function, F(alpha). The more powerful the test is in detecting initial bias the more rapidly F(alpha) should go from 0 to 1 as alpha increases from 0 to 1, when initial bias is present. Results: Schruben's Max test (20) performed well on all models (model4, M/M/1, being the worst). Test appeared valid with reasonable power in detecting bias. Care needs to be taken to test for the correct sign (i.e. positive or negative bias). Test results indicates that the asymptotic theory that the test is based on produces good results. Many of the simulation runs here were quite short & results therefore improved for longer runs. Test most useful when transient mean function is almost constant or dissipates rapidly. Test not powerful when a large initial transient is present that remains throughout run. The Area test (26) is far better in that situation. Ockerman & Goldsman (1999) [22]: Systems Studied: Model1: A step process Y(j), j=1,2,... where Y(j)s = ±1 with probability 0.5. The variance is 1. Model2: AR(1) with Normal errors. The variance is 19. Model3: MA(1) with Normal errors. The variance is 0.01. Model4: M/M/1 where output variable is waiting time. The arrival rate is 0.8, service rate is 1 and the variance is stated as being 1976?? Model5: 2-station re-entrant line, with a poisson arrival process. Performance criteria: P{Reject H0 | H0 is true) and P{Reject H0 | H0 is false}. Results: P{Reject H0 | H0 is true) criteria: All methods performed relatively well (reaching 10% value with increasing n). P{Reject H0 | H0 is false} criteria: The Ockerman & Goldsman Students t-tests Method (35), Ockerman & Goldsman (t-test) Compound Tests (36) and Schruben max test (20) seem to cope a little better than the other variance ratio tests BM Batch Means Test (25) and Area test (26) when the M/M/1 or re-entrant line systems started empty. That is, they cope with "2nd-order effect" slightly better. Yucesan (1993) [23]: Randomisation test (19) generally more conservative than Conway Rule (10) or (graphical) Schruben (20).

	Law (1983) [34]: No tests done - just algorithm explained & commented on: "Schruben tested his procedure in a large number of independent experiments for each of a number of stochastic models…test generally had high power…..Performed well for fairly small values of n though based on asymptotic properties". Law & Kelton(2000) [52]: Brief explanation of Schruben's procedures. "Schruben tested his procedure on several stochastic models with a known value of [steady state mean], and found that it had high power in detecting initialization bias."
Positives	"The computation involved … was minimal even on the smallest of computers" [18]. Performed well on all models. Test appeared valid with reasonable power in detecting initial bias. Test most useful when transient mean fn is almost constant or dissipates rapidly. [20] Performed well for fairly small values of n (though based on asymptotic properties [34] [20]. Schruben max test (20) seemed to cope a little better than other variance ratio tests: BM Batch means test (25) & BM Area test (26); when M/M/1 or re-entrant line systems started empty - i.e coped better with "2nd-order effect". [22]
Criticisms	Schruben max test (20) performed poorly for 2nd model (M/M/1 $\rho = 0.9$) [18]. Worst performance for model4, M/M/1. Care needs to be taken to test for the correct sign (i.e positive or negative bias). Test not powerful when a large initial transient is present that remains throughout run. [20]
Accuracy	Medium
Simplicity	Medium
Automation potential	Medium – need to access whether bias is +ve or –ve.
Generality (No assumptions)	Good/Medium – asymptotic assumption seems robust
Parameters to estimate?	3 – run length (m); variance (V); degree of freedom (δ) for F dist test. (+ number of replications if used)
Computation time	Good

Method ID	21
Method Name	Schruben's Modified Test
Brief description	Nelson (1982) [16]: Null hypothesis, H0: no negative initial condition bias in mean of the output process.

	"The output of a single replication is divided in half. If the run is long enough then any bias is most prevalent in the first half. The cusum values from each half are compared in terms of the location and magnitude of their maximum deviation from zero. If the behaviour of the first half is significantly different from the second half, then the hypothesis of no initial condition bias is rejected. There are several versions of this test that have power against specific alternatives." Preston White Jnr (2000) [28]: Schruben recommends batching parameters be set at b, number of batches = Int[n / 5]
Literature Review	Nelson (1982) [16]: No tests done - just algorithm explained & commented on: "version given here is conservative in that it tries to ensure that all of the asymptotic assumptions behind the test will be valid" Preston White Jnr et al (2000) [28]: Systems Studied: 2nd order autoregressive process with zero mean and with differing initial parameter values. 3 bias fns: exponential, mean shift, under-damped oscillations (based on Cash et al (1992) [33]). Bias fns incorporated by superposition (adding into output) and injection (adding into state equation). Performance Criteria: Sample mean, abs bias = |grand estimated mean - grand mean of unbiased data|; p-value of 2 sample t-test - H0: estimated mean = mean from unbiased data; average computation time; min, max, mean & standard deviation of truncation pt.; number of inconclusive results. Methods tested: MCR (14), MSER-5 (15), Schruben's modified test (21), BM Max test (24), & BM Batch means test (25). The MCR/MSER method is applied sequentially. Results: Schruben's Modified Test (21) is effective in removing damped oscillating bias and mean shift bias. Generally not successful in dealing with exponential bias except in most extreme cases. Occasionally removed large amount of data relative to other methods tested. Not effective in detecting bias in capacitated damped oscillating bias. Law (1983) [34]: No tests done - just algorithm explained & commented on: "May be a more powerful test than Schruben's original test (20), for large initial bias" Law & Kelton (2000) [52]: Brief explanation of Schruben's procedures. "Schruben tested his procedure on several stochastic models with a known value of [steady state mean], and found that it had high power in detecting initialization bias."

Positives	Effective in removing damped oscillating bias and mean shift bias. [28] "May be a more powerful test than Schruben's original test (20), for large initial bias" [34] High power in detecting initialization bias [52].
Criticisms	Generally not successful in dealing with exponential bias except in most extreme cases. Not effective in detecting bias in capacitated damped oscillating bias. Occasionally removed large amount of data relative to other methods MCR (14), MSER-5 (15), BM Max test (24), & BM Batch means test (25). [28] "version given [in Nelson (1982) [16]] … is conservative in that it tries to ensure that all of the asymptotic assumptions behind the test will be valid"
Accuracy	Medium
Simplicity	Medium
Automation potential	Medium - need to access whether bias is +ve or –ve.
Generality (No assumptions)	Good/ Medium – need to either decide if bias is +ve or –ve
Parameters to estimate?	4 – run length (n); variance (V); degree of freedom (δ) for F dist test and Batch size (b)
Computation time	Good

Method ID	22
Method Name	Optimal Test (Brownian bridge process)
Brief description	Schruben et al (1983) [18]: Replicate k times with a run length of n. Compute the variance σ^2 and degrees of freedom d. Compute the test statistic: $\hat{T} = \left(\sqrt{45} / n^{3/2} \hat{\sigma}\right) \sum_{k=1}^{n} \left(1 - k/n\right) k \left(\bar{Y}_n - \bar{Y}_k\right)$, where n is the number of observations in time series, and $\bar{Y}_k = \frac{1}{k} \sum_{i=1}^{k} Y(i)$, where Y(i) is the value of the i^{th} observation for each of the k replications. Reject the null hypothesis is no negative initial bias if T > t(d,α). Here t(d,α) is the upper 100α quantile of the t distribution with d degrees of freedom.

	If a test for positive bias is required, then the sign of the T statistic should be changed. The absolute value of T can be used if a two-sided t test is required.
Literature Review	Pawlikowski(1990) [7]: "Despite the sophisticated theory hidden behind these tests, they appear to be simple numerically and can be applied to a wide class of simulated processes"…"The main practical problem with their implementation is that they require a priori knowledge of the variance estimator of the simulated process in steady state." Schruben et al(1983) [18]: Systems Studied: Model1: AR(2) with oscillating initial transient function. Model2: M/M/1 queue with $\rho = 0.9$. Output variable is waiting times. Model3: Simple production-inventory model. Model4: time-shared computer system (Adiri & Avi-Itzhak [1969]). Model5: Telephone exchange model (M/M/s). Output is number of busy lines per time period. Model6: Network of 3 capacitated M/M/s queues with feedback. Performance criteria: Percentage of runs for which the null hypothesis of no bias was rejected. Results: The Optimal Test (22) had high power in detecting presence of initialisation bias in all test models and compared well with Schruben's Max Test (20). Schruben's Max Test (20) performed poorly for 2nd model (M/M/1 $\rho = 0.9$). "The computation involved in the two tests was minimal even on the smallest of computers". "The Optimal Test (22) for initialisation bias in simulations presented in this paper appears to be robust and powerful in widely differing situations. The forms of the initial transient mean function in the models used in the examples are quite different. Relatively short runs were used in the experiments to illustrate the robustness of the asymptotic theory on which the test is based." "We recommend that both the test presented here, [Optimal Test (22)] and the test presented [by Schruben, Schruben's Max Test (20)] be conducted to support ascertations that initialisation bias in the mean of a simulation output process has been effectively controlled." Ma & Kochner(1993) [31]: Systems Studied: Artificial stochastic sequences are produced, Y(i) = X(i) +u(i), i=1,2,…,n; where X(i) is a series of i.i.d N(0,1) random variables & u(i) is the initial bias function, a geometrically declining function

	which converges to 0 as i increases: $u(i) = u(1)\, p^{(i-1)}$, $(0<p<1)$; u(1) & p are predetermined parameters. (Also possible to use AR(1) model as X, although not clear if done here). Performance Criteria: The number of positive bias sequences detected by test for all 6 combinations of u and p parameters. Results: Optimal Test (22) more sensitive to bias than Rank test (23). Sample size affects performance of each test: too large or too small a sample size reduces power of each test. Kimbler & Knight (1987) [46]: Systems Studied: M/M/1 queue with $\rho = 0.9$. This is a "highly congested system …has an unusually gradual transition into steady state" Performance Criteria: Steady state statistics (time in system) compared with theoretical solution using Dudewicz method for determining the best of k systems. Results: Though quite complex to (code) execute, the Optimal Test (22) worked well. "…We also found that while the methods proposed by both Kelton & Law (18) and Schruben [Optimal Test (22)] were very complex and sophisticated, the results which they produced did not warrant the additional level of difficulty". All methods tested: Relaxation Heuristics (17); Kelton & Law Regression (18); Optimal Test (22) & Kimbler's double exponential smoothing (43), worked equally well in relation to estimating mean time in system...but there was quite a difference in truncation points. The Relaxation Heuristics (17) truncated far later than others; then Kelton & Law Regression (18) and Optimal Test (22), with Kimbler's double exponential smoothing (43) being the least conservative. Relaxation Heuristics (17) and Kimbler's double exponential smoothing (43) are far simpler than Kelton & Law Regression (18) and Optimal Test (22) to carry out. Law & Kelton (2000) [52]: Only a brief mention (with cite) of this initialisation bias test as a variation of the Schruben procedures.
Positives	Simple numerically. Applicable to a wide class of simulated processes [7]. High power in detecting presence of initialisation bias. "Compared well with Schruben's Max Test (20)." "Robust and powerful in widely differing situations." Asymptotic theory robust [18] Optimal Test (22) more sensitive to bias than Rank test (23). [31]

Criticisms	Require estimate of Variance [7]. Sample size affects performance of test: too large or too small a sample size reduces power of test. [31] Quite complex to (code) execute. [46]
Accuracy	Medium/Good
Simplicity	Medium
Automation potential	Good/Medium – need to either decide if bias is +ve or –ve or use the two-sided test.
Generality (No assumptions)	Good/Medium – Asymptotic assumptions a possible problem – need to either decide if bias is +ve or –ve or use the two-sided test.
Parameters to estimate?	4-5 – Variance (V), degrees of freedom (d), run length (n), number of reps (k). May need to use batches (b)
Computation time	Good

Method ID	23
Method Name	Rank Test
Brief description	Vassilacopoulos (1989) [14]: Let $x_1, x_2, \ldots, x_N$ be the observations to be tested for initial bias (these can be from a single run or averages from several runs etc..). Let R_k be the ranking for observation x_k, $_{(k=1,2,\ldots,N)}$, where $R_k = \sum_{i=1}^{N} f(x_k - x_i)$, & where $f(x) = \begin{cases} -1, x < 0 \\ 0, x = 0 \\ 1, x > 0 \end{cases}$ Let $W_k = \sum_{j=1}^{k} R_j$ and $\mathbf{U_N(k) = 2W_k - k(N+1)}$, k=1,2,…,N; $U_N(0)=0$ Let the test statistic be: $$c = \max_k \lvert U_N(k) \rvert$$ and the significance probabilities be $$\hat{\alpha}^{+} = \exp\left\{ -6c^2 \big/ \left(N^3 + N^2\right) \right\}$$, likewise for $\hat{\alpha}^{-}$;

	$$\hat{\alpha} \cong 2\exp\left\{-6c^2 \Big/ \left(N^3 + N^2\right)\right\}$$ 1. Find the ranks (R_k) of the observations and calculate $\{U_N(k);\ \ k=1,2,\ldots,N\}$: 2. Find c and the significance probability; $\hat{\alpha}$(for two-sided test); $\hat{\alpha}^+$ (for +ve bias); $\hat{\alpha}^-$ (for –ve bias); associated with it. 3. Reject the hypothesis of "no initial bias" if $\hat{\alpha} < \alpha$ where α is the selected significance level of the test.
Literature Review	Vassilacopoulos (1989) [14]: Systems Studied: M/M/1 & M/M/4 queues with $\rho = 0.9$. There are 3 initial states used: empty & idle (producing negative bias), equal to steady state mean (producing no bias), twice steady state mean (producing positive bias). The output variable is customer waiting times. Results: Method performed well, detecting bias with high power. Ma & Kochner(1993) [31]: Systems Studied: Artificial stochastic sequences are produced, Y(i) = X(i) +u(i), i=1,2,…,n; where X(i) is a series of i.i.d N(0,1) random variables & u(i) is the initial bias function, a geometrically declining function which converges to 0 as i increases: $u(i) = u(1)\, p^{(i-1)}$, (0<p<1); u(1) & p are predetermined parameters. (Also possible to use AR(1) model as X, although not clear if done here). Performance Criteria: The number of positive bias sequences detected by test for all 6 combinations of u and p parameters. Results: Optimal Test (22) more sensitive to bias than Rank test (23). Sample size affects performance of each test: too large or too small a sample size reduces power of each test. Law & Kelton(2000) [52]: Only a brief mention (with cite) of this initialisation bias test. "Limited testing on the M/M/s queue produced encouraging results"
Positives	Performed well, detecting bias with high power. [14]
Criticisms	Less sensitive to bias than Optimal Test (22). Sample size affects performance of test: too large or too small a sample size reduces power of test. [31]
Accuracy	Medium/Good

Simplicity	Medium
Automation potential	Good
Generality (No assumptions)	Good – asymptotic theory seems robust
Parameters to estimate?	1-2 – run length (N) and number of replications if used.
Computation time	Medium

Method ID	24
Method Name	Batch means based test – Max test
Brief description	Cash et al (1992) [33]: "…a test based on the location and magnitude of the maximum deviation…" For a test for negative bias (positive bias test requires change of sign for test statistic): Partition $X_1, X_2, \ldots, X_n$ into b non-overlapping batches of m observations such that n = bm. Let, $\bar{X}_{i,j} = \frac{1}{j}\sum_{t=1}^{j} X_{(i-1)m+t}$, be the cumulative within-batch means for i=1,2,…,b and j=1,2,…,m. $S_{i,j} = \bar{X}_{i,m} - \bar{X}_{i,j}$ $\hat{K}_i = \arg\max_{1 \le j \le m}\{jS_{i,j}\}$ $\hat{S}_i = \hat{K}_i S_{i,\hat{K}_i}$ $Q_{MAX} = \sum_{i=1}^{b} \frac{m\hat{S}_i^2}{\hat{K}_i(m-\hat{K}_i)}$ And $V_{BM+MAX} = \frac{Q_{BM+MAX}}{4b-1}$, be the batch-means variance estimator, where $Q_{BM+MAX} = Q_{BM} + Q_{MAX}$, for definition of Q_{BM} see Method [25].

	Partition the b batch means into two groups consisting of the first b' batch means and the last (b-b') batch means. Let V_{BM+MAX}^{1st} be the variance estimator calculated using only the first b' batches; and let V_{BM+MAX}^{2nd} be the variance estimator calculated using the last (b-b') batches only. Under the null hypothesis of no initial bias, the ratio $F_{BM+MAX} = V_{BM+MAX}^{1st} / V_{BM+MAX}^{2nd}$ converges in distribution to an F random variable. The critical value for the test is $F_{4b'-1, 4b-4b'-1, (1-\alpha)}$.
Literature Review	Preston White Jnr et al (2000) [28]: Systems Studied: 2nd order autoregressive process with zero mean and with differing initial parameter values. 3 bias fns: exponential, mean shift, under-damped oscillations (based on Cash et al (1992) [33]). Bias fns incorporated by superposition (adding into output) and injection (adding into state equation). Performance Criteria: Sample mean, abs bias = |grand estimated mean - grand mean of unbiased data|; p-value of 2 sample t-test - H0: estimated mean = mean from unbiased data; average computation time; min, max, mean & standard deviation of truncation pt.; number of inconclusive results. Methods tested: MCR (14), MSER-5 (15), Schruben's modified test (21), BM Max test (24), & BM Batch means test (25). The MCR/MSER method is applied sequentially. Results: Principle shortcoming of the BM Max test (24) was that it was often inconclusive particularly for initial parameters >0. Did perform well on data sets containing damped oscillating bias. Generally efficient in picking truncation pts. Not effective in detecting bias in capacitated damped oscillating bias. Generally unreliable because the test statistic on which it is based is frequently undefined yielding an inconclusive result. "Must test for both positive and negative bias at each iteration". Cash et al (1992) [33]: Systems Studied: Model1: AR(1). Model2: M/M/1 queue, $\rho = 0.5, 0.8$. Output variable is 'delay in queue per customer'. Model3: Markov Chain. Bias fns: (1) Mean shift bias; (2) Linear bias fn; (3) Quadratic bias fn (based on GSS) (4) Damped

	oscillating bias function. All 4 biases are made to go to zero at a fixed n (data number). Uses (and recommends using) batch number, b, ≤ 16. Performance Criteria: Bias Index: Absolute Bias of point estimator / (asymptotic variance of process / n). Estimated power of test. P(test rejects H0 of no bias| no bias present). Methods tested: BM Max test (24), BM Batch means test (25), BM Area test (26). Results: Power of tests are reduced by making the number of batches (b) too small or too big (But of course the practitioner doesn't know what is "too small" or "too big"!). The slower the bias decays the more difficulty the tests have in detecting it. BM Max test (24) is the most powerful test of the ones tested here. Recommend Max test run with 8 batches. More work on a good deletion strategy still needed. Lee & Oh (1994) [42]: Systems Studied: M/M/1/infinity queuing model with ρ = 0.5; 0.7; 0.9. Output variable is waiting time in queue. Performance criteria: Effectiveness was assessed using a bias calculation: Abs[true mean - estimated mean] / estimated mean Results: Methods M1 (40) and BM Max test (24) "detect a truncation pt regardless of system state". But M2 (40) "can determine correctly the transient state…" Goldsman et al (1994) [43]: Systems Studied: AR(1)~Norm error; AR(1)~exp error; M/M/1 queue. Performance criteria: Analytical power results if possible or empirical power results. Methods tested: BM Max test (24), BM Batch means test (25), BM Area test (26) Results: Tests valid for all examples. ". tests fared particularly well when [number of batches and number of batches in first group] were large. Unfortunately when we divided simulation output into many batches the performance of the non-BM tests was disappointing. …A conservative yet powerful approach may be to use the area or maximum estimators after dividing the simulation into a smaller number of batches…" Law & Kelton(2000) [52]: Only a brief mention (with cite) of this initialisation bias test as a variation of the

	Schruben procedures.
Positives	Performed well on data sets containing damped oscillating bias (except capacitated ones) [28] Tests fared particularly well when [number of batches and number of batches in first group] were large. [43]
Criticisms	Often inconclusive - Generally unreliable because the test stat on which it is based is frequently undefined yielding an inconclusive result. Not effective in detecting bias in capacitated damped oscillating bias. "Must test for both +ve and -ve bias at each iteration". [28] The slower the bias decays the more difficulty the tests have in detecting it. BM Max test (24) is more powerful than BM Batch means test (25) and BM Area test (26). [33] Method will "detect a truncation pt regardless of system state". [42]
Accuracy	Poor/Medium
Simplicity	Medium
Automation potential	Medium – need to access whether bias is +ve or –ve.
Generality (No assumptions)	Good/Medium – asymptotic assumption seems robust
Parameters to estimate?	4 - Run length (n), batch number (b), division of batches into two groups (b'), variance (V)
Computation time	Good

Method ID	25
Method Name	Batch means based tests – Batch Means Test
Brief description	Partition $X_1, X_2, \ldots, X_n$ into b nonoverlapping batches of m observations such that n=bm. Let, $\overline{X}_i = \frac{1}{m}\sum_{j=1}^{m} X_{(i-1)m+j}$, be the batch means for i=1,2,…,b. And $V_{BM} = \frac{Q_{BM}}{b-1}$, be the batch-means variance estimator, where $Q_{BM} = m\sum_{i=1}^{b}\left[\overline{X}_i - \frac{1}{b}\sum_{j=1}^{b}\overline{X}_j\right]^2$ Partition the b batch means into two groups consisting of the first b' batch means and the last (b-b') batch means.

	Let V_{BM}^{1st} be the variance estimator calculated using only the first b' batches; and let V_{BM}^{2nd} be the variance estimator calculated using the last (b-b') batches only. Under the null hypothesis of no initial bias, the ratio $F_{BM} = V_{BM}^{1st} / V_{BM}^{2nd}$ converges in distribution to an F random variable. The critical value for the test is $F_{b'-1, b-b'-1, (1-\alpha)}$.
Literature Review	Ockerman & Goldsman (1999) [22]: Systems Studied: Model1: A step process Y(j), j=1,2,... where Y(j)s = ±1 with probability 0.5. The variance is 1. Model2: AR(1) with Normal errors. The variance is 19. Model3: MA(1) with Normal errors. The variance is 0.01. Model4: M/M/1 where output variable is waiting time. The arrival rate is 0.8, service rate is 1 and the variance is stated as being 1976?? Model5: 2-station re-entrant line, with a poisson arrival process. Performance criteria: P{Reject H0 | H0 is true) and P{Reject H0 | H0 is false}. Results: P{Reject H0 | H0 is true) criteria: All methods performed relatively well (reaching 10% value with increasing n). P{Reject H0 | H0 is false} criteria: The Ockerman & Goldsman Students t-tests Method (35), Ockerman & Goldsman (t-test) Compound Tests (36) and Schruben max test (20) seem to cope a little better than the other variance ratio tests BM Batch Means Test (25) and Area test (26) when the M/M/1 or re-entrant line systems started empty. That is, they cope with "2nd-order effect" slightly better. Preston White Jnr et al (2000) [28]: Systems Studied: 2nd order autoregressive process with zero mean and with differing initial parameter values. 3 bias fns: exponential, mean shift, under-damped oscillations (based on Cash et al (1992) [33]). Bias fns incorporated by superposition (adding into output) and injection (adding into state equation). Performance Criteria: Sample mean, abs bias = |grand estimated mean - grand mean of unbiased data|; p-value of 2 sample t-test - H0: estimated mean = mean from unbiased data; average computation time; min, max, mean & standard deviation of truncation pt.; number of inconclusive results. Methods tested: MCR (14), MSER-5 (15), Schruben's modified test (21), BM Max test (24), & BM Batch

	means test (25). The MCR/MSER method is applied sequentially. Results: BM Batch means test (25) is the worst least effective method tested. It had relatively low sensitivity in detecting bias and rarely truncated enough of series. It never, however, truncated beyond the optimal truncation pt and was generally fast to run. Cash et al (1992) [33]: Systems Studied: Model1: AR(1). Model2: M/M/1 queue, ρ = 0.5, 0.8. Output variable is 'delay in queue per customer'. Model3: Markov Chain. Bias fns: (1) Mean shift bias; (2) Linear bias fn; (3) Quadratic bias fn (based on GSS) (4) Damped oscillating bias function. All 4 biases are made to go to zero at a fixed n (data number). Uses (and recommends using) batch number, b, ≤ 16. Performance Criteria: Bias Index: Absolute Bias of point estimator / (asymptotic variance of process / n). Estimated power of test. P(test rejects H0 of no bias| no bias present). Methods tested: BM Max test (24), BM Batch means test (25), BM Area test (26). Results: Power of tests are reduced by making the number of batches (b) too small or too big (But of course the practitioner doesn't know what is "too small" or "too big"!). The slower the bias decays the more difficulty the tests have in detecting it. BM Batch means test (25) & BM Area test (26) are joint least powerful tests of the ones tested here. More work on a good deletion strategy still needed. Goldsman et al (1994) [43]: Systems Studied: AR(1)~Norm error; AR(1)~exp error; M/M/1 queue. Performance criteria: Analytical power results if possible or empirical power results. Methods tested: BM Max test (24), BM Batch means test (25), BM Area test (26) Results: Tests valid for all examples. All tests fared particularly well when [number of batches and number of batches in first group] were large. Unfortunately when we divided simulation output into many batches the performance of the non-BM tests was disappointing. "…A conservative yet powerful approach may be to use the area or maximum estimators after dividing the simulation into a smaller number of batches…"

	Law & Kelton(2000) [52]: Only a brief mention (with cite) of this initialisation bias test as a variation of the Schruben procedures.
Positives	P{Reject H0 | H0 is true) criteria: performed relatively well (reaching 10% value with increasing n) [22]. Never truncated beyond the optimal truncation pt. Generally fast to run. [28] Fared particularly well when [number of batches and number of batches in first group] were large. [43]
Criticisms	Worst, least effective method when compared with methods MCR (14), MSER-5 (15), Schruben's modified test (21) and BM Max test (24). Had relatively low sensitivity in detecting bias and rarely truncated enough of series [28]. Power of test is reduced by making the number of batches (b) too small or too big. The slower the bias decays the more difficulty the test has in detecting it. Less powerful test than BM Max test (24). [33]
Accuracy	Poor/Medium
Simplicity	Medium
Automation potential	Good
Generality (No assumptions)	Good – though based on asymptotic theory
Parameters to estimate?	4 - Run length (n), batch number (b), division of batches into two groups (b'), variance (V)
Computation time	Good

Method ID	26
Method Name	Batch means based tests – Area Test
Brief description	Cash et al (1992) [33]: A test based on the area under a standardised time series… Partition $X_1,X_2,\ldots,X_n$ into b nonoverlapping batches of m observations such that n=bm. Transform the data into b standardised time series and compute a variance estimator based on the area under the standardised time series, as follows, where i=1,2,…,b and j=1,2,…,m and $0 \le t \le 1$. Let,

	$\bar{X}_{i,j} = \frac{1}{j}\sum_{t=1}^{j} X_{(i-1)m+t}$, be the cumulative within-batch means $T_{i,m}(t) = \frac{\lfloor mt \rfloor \left(\bar{X}_{i,m} - \bar{X}_{i,\lfloor mt \rfloor}\right)}{\sigma\sqrt{m}}$ $\hat{A}_i = \frac{\sigma}{m}\sum_{j=1}^{m}\left[\sqrt{12}T_{i,m}(j/m)\right]$ $Q_{AREA} = \sum_{i=1}^{b}\hat{A}_i^{2}$ where σ^2 is the asymptotic variance constant of the output process. Let $V_{BM+AREA} = \frac{Q_{BM+AREA}}{2b-1}$, be the batch-means variance estimator, where $Q_{BM+AREA} = Q_{BM} + Q_{AREA}$, for definition of Q_{BM} see Method (25). Partition the b batch means into two groups consisting of the first b' batch means and the last (b-b') batch means. Let $V_{BM+AREA}^{1st}$ be the variance estimator calculated using only the first b' batches; and let $V_{BM+AREA}^{2nd}$ be the variance estimator calculated using the last (b-b') batches only. Under the null hypothesis of no initial bias, the ratio $F_{BM+AREA} = V_{BM+AREA}^{1st} / V_{BM+AREA}^{2nd}$ converges in distribution to an F random variable. The critical value for the test is $F_{2b'-1, 2b-2b'-1, (1-\alpha)}$.
Literature Review	Ockerman & Goldsman (1999) [22]: Systems Studied: Model1: A step process Y(j), j=1,2,... where Y(j)s = ±1 with probability 0.5. The variance is 1.

	Model2: AR(1) with Normal errors. The variance is 19. Model3: MA(1) with Normal errors. The variance is 0.01. Model4: M/M/1 where output variable is waiting time. The arrival rate is 0.8, service rate is 1 and the variance is stated as being 1976?? Model5: 2-station re-entrant line, with a poisson arrival process. Performance criteria: P{Reject H0 | H0 is true) and P{Reject H0 | H0 is false}. Results: P{Reject H0 | H0 is true) criteria: All methods performed relatively well (reaching 10% value with increasing n). P{Reject H0 | H0 is false} criteria: The Ockerman & Goldsman Students t-tests Method (35), Ockerman & Goldsman (t-test) Compound Tests (36) and Schruben max test (20) seem to cope a little better than the other variance ratio tests BM Batch Means Test (25) and Area test (26) when the M/M/1 or re-entrant line systems started empty. That is, they cope with "2nd-order effect" slightly better. Cash et al (1992) [33]: Systems Studied: Model1: AR(1). Model2: M/M/1 queue, $\rho = 0.5, 0.8$. Output variable is 'delay in queue per customer'. Model3: Markov Chain. Bias fns: (1) Mean shift bias; (2) Linear bias fn; (3) Quadratic bias fn (based on GSS) (4) Damped oscillating bias function. All 4 biases are made to go to zero at a fixed n (data number). Uses (and recommends using) batch number, b, ≤ 16. Performance Criteria: Bias Index: Absolute Bias of point estimator / (asymptotic variance of process / n). Estimated power of test. P(test rejects H0 of no bias| no bias present). Methods tested: BM Max test (24), BM Batch means test (25), BM Area test (26). Results: Power of tests are reduced by making the number of batches (b) too small or too big (But of course the practitioner doesn't know what is "too small" or "too big"!). The slower the bias decays the more difficulty the tests have in detecting it. BM Batch means test (25) & BM Area test (26) are joint least powerful tests of the ones tested here. More work on a good deletion strategy still needed.

	Goldsman et al (1994) [43]: Systems Studied: AR(1)~Norm error; AR(1)~exp error; M/M/1 queue. Performance criteria: Analytical power results if possible or empirical power results. Methods tested: BM Max test (24), BM Batch means test (25), BM Area test (26) Results: Tests valid for all examples. All tests fared particularly well when [number of batches and number of batches in first group] were large. A conservative yet powerful approach may be to use the area or maximum estimators after dividing the simulation into a smaller number of batches…" Law & Kelton(2000) [52]: Only a brief mention (with cite) of this initialisation bias test as a variation of the Schruben procedures.
Positives	P{Reject H0 | H0 is true) criteria: performed relatively well (reaching 10% value with increasing n). [22] Fared particularly well when [number of batches and number of batches in first group] were large. [43]
Criticisms	Power of test is reduced by making the number of batches (b) too small or too big. The slower the bias decays the more difficulty the test has in detecting it. Less powerful than BM Max test (24). [33]
Accuracy	Poor/Medium
Simplicity	Medium
Automation potential	Good
Generality (No assumptions)	Good – based on asymptotic theory
Parameters to estimate?	5 - Run length (n), batch number (b), division of batches into two groups (b'), Standardisation parameter (t), asymptotic variance of output process (σ^2) and variance of area under standardised time series (V)
Computation time	Good

Method ID	35
Method Name	Ockerman & Goldsman Students t-tests Method
Brief description	Find the sample means of the first and second halves of the data series: $\bar{Y}_{2,n/2}$ and $\bar{Y}_{2,n/2}$ Create an asymptotically independent χ^2 estimate of the variance of $\bar{Y}_{2,n/2}$-$\bar{Y}_{2,n/2}$, namely $V_{BM,8}^{(2)}(n)$ (variance of 2nd half of data series using 8 batches).

	Construct the statistic: $Q_{BM} = \dfrac{\bar{Y}_{2,n/2} - \bar{Y}_{1,n/2}}{\sqrt{(4/n)V^{(2)}_{BM,8}(n)}}$ Which converges to a t-dist random variable, with 7 df, under the H_0 of no initialisation bias. i.e. this test rejects H_0 at the 100α% level if $Q_{BM} > t_{7,\alpha}$ is the upper-α quantile from the t-dist with 7 df. NOTE: this test is one-sided based on the assumption of H_1: negative initial bias present. Multiply test statistic by -1 to test for positive bias.
Literature Review	Ockerman & Goldsman (1999) [22]: Systems Studied: Model1: A step process Y(j), j=1,2,... where Y(j)s = ±1 with probability 0.5. The variance is 1. Model2: AR(1) with Normal errors. The variance is 19. Model3: MA(1) with Normal errors. The variance is 0.01. Model4: M/M/1 where output variable is waiting time. The arrival rate is 0.8, service rate is 1 and the variance is stated as being 1976?? Model5: 2-station re-entrant line, with a poisson arrival process. Performance criteria: P{Reject H0 | H0 is true) and P{Reject H0 | H0 is false}. Results: P{Reject H0 | H0 is true) criteria: All methods performed relatively well (reaching 10% value with increasing n). P{Reject H0 | H0 is false} criteria: The Ockerman & Goldsman Students t-tests Method (35), Ockerman & Goldsman (t-test) Compound Tests (36) and Schruben max test (20) seem to cope a little better than the other variance ratio tests BM Batch Means Test (25) and Area test (26) when the M/M/1 or re-entrant line systems started empty. That is, they cope with "2nd-order effect" slightly better. Results generally good. "It would appear that for stochastic processes that have boundaries (e.g. the number of customers waiting in a queue cannot be negative), if the process is started near one of the boundaries and the variance at the boundary is lower than the variance at steady state, then many current initialisation bias tests will not be successful in detecting the bias. We believe that the t-test method (35) and compound test (36) are steps towards a solution." Schruben max test (20) and new tests Ockerman & Goldsman Students t-tests Method (35), Ockerman &

	Goldsman (t-test) Compound Tests (36) slight improvement on BM Batch Means Test (25) & BM Area test (26).
Positives	P{Reject H0 | H0 is true) criteria: performed relatively well (reaching 10% value with increasing n). P{Reject H0 | H0 is false} criteria: Schruben max test (20) seemed to cope a little better than other variance ratio tests, BM Batch Means Test (25) and Area test (26) when systems (M/M/1 or re-entrant line) started empty - i.e cope with "2nd-order effect" slightly better. [22]
Criticisms	Need to know whether bias is positive of negative and adjust statistic accordingly.
Accuracy	Medium
Simplicity	Medium
Automation potential	Medium - Need to know whether bias is positive of negative and adjust statistic accordingly.
Generality (No assumptions)	Good/Medium - Need to know whether bias is positive of negative and adjust statistic accordingly.
Parameters to estimate?	3 - Run length (n), variance (V), sig level (α),
Computation time	Good

Method ID	36
Method Name	Ockerman & Goldsman (t-tests) Compound Methods
Brief description	Component1: t-test method(as in method 35) but using the batched-area estimator for the variance so that independence between this test and the one in component2 is maintained: $V_{BM,8}^{(2)}(n) = \frac{1}{8}\sum_{i=1}^{8} V_{1,1}[i]$ i.e. the average of 8 area estimators from the 2^{nd} half of the data series. $Q_{BM} = \frac{\bar{Y}_{2,n/2} - \bar{Y}_{1,n/2}}{\sqrt{(4/n)V_{BM,8}^{(2)}(n)}}$ now converges to the t-dist with 8df under the null hypothesis of no initial bias. i.e. this test rejects H_0 at the 100α% level if $Q_{BM} > t_{8,\alpha}$ is the upper-α quantile from the t-dist with 8 df. Component2: Method (25), the Batch means test , using 8 batches in each half of the data series. --

	Compound Test 1: Reject the claim of no initial bias only if BOTH the t-test and the batch-means test would have individually rejected the claim. Individual tests constructed with size $\sqrt{\alpha}$ to obtain a compound test with size α. Compound Test 2: Reject the claim of no initial bias only if EITHER the t-test or the batch-means test would have individually rejected the claim. Individual tests constructed with size $1-\sqrt{(1-\alpha)}$ to obtain a compound test with size α.
Literature Review	Ockerman & Goldsman (1999) [22]: Systems Studied: Model1: A step process Y(j), j=1,2,... where Y(j)s = ±1 with probability 0.5. The variance is 1. Model2: AR(1) with Normal errors. The variance is 19. Model3: MA(1) with Normal errors. The variance is 0.01. Model4: M/M/1 where output variable is waiting time. The arrival rate is 0.8, service rate is 1 and the variance is stated as being 1976?? Model5: 2-station re-entrant line, with a poisson arrival process. Performance criteria: P{Reject H0 | H0 is true) and P{Reject H0 | H0 is false}. Results: P{Reject H0 | H0 is true) criteria: All methods performed relatively well (reaching 10% value with increasing n). P{Reject H0 | H0 is false} criteria: The Ockerman & Goldsman Students t-tests Method (35), Ockerman & Goldsman (t-test) Compound Tests (36) and Schruben max test (20) seem to cope a little better than the other variance ratio tests BM Batch Means Test (25) and Area test (26) when the M/M/1 or re-entrant line systems started empty. That is, they cope with "2nd-order effect" slightly better. Results generally good. Compound2 probably better than Compound1. "It would appear that for stochastic processes that have boundaries (e.g. the number of customers waiting in a queue cannot be negative), if the process is started near one of the boundaries and the variance at the boundary is lower than the variance at steady state, then many current initialisation bias tests will not be successful in detecting the bias. We believe that the Ockerman & Goldsman Students t-tests Method (35) and Ockerman & Goldsman (t-test) Compound Tests (36) are steps towards a solution."

	Schruben max test (20) and new tests (35) & (36) slight improvement on BM Batch Means Test (25) and Area test (26).
Positives	P{Reject H0 | H0 is true) criteria: performed relatively well (reaching 10% value with increasing n). P{Reject H0 | H0 is false} criteria: Schruben max test (20) seemed to cope a little better than other variance ratio tests, BM Batch Means Test (25) and Area test (26) when systems (M/M/1 or re-entrant line) started empty - i.e cope with "2nd-order effect" slightly better. [22]
Criticisms	Need to know whether bias is positive of negative and adjust statistic accordingly.
Accuracy	Medium
Simplicity	Medium
Automation potential	Medium - Need to know whether bias is positive of negative and adjust statistic accordingly.
Generality (No assumptions)	Good/Medium - Need to know whether bias is positive of negative and adjust statistic accordingly.
Parameters to estimate?	4 - Run length (n), two different variances (V_{BM}) & (V_{BA}), sig level (α),
Computation time	Good/Medium

(Return to top of document)

STATISTICAL METHODS

Method ID	16
Method Name	Goodness-Of-Fit Test
Brief description	Pawlikowski (1990) [7]: "...the sequence of observations should be partitioned into batches of at least m_0 = 10 observations each; e.g. Soloman (1983) selected m_0 = 30… In a time series of observations $x_1, x_2, \ldots, x_n$ the initial transient is over after n_0 observations if the χ^2 goodness-of-fit test confirms that in the batch of observations $x_{n_0+1}, x_{n_0+2}, \ldots, x_{n_0+m_0}$ following the observation n_0, the numbers of observations above and below the running mean $\overline{X}(n_0)$ are about the same."
Literature Review	Pawlikowski (1990) [7]: "approach, which Solomon (1983) [*"Simulation of Waiting-Line Systems"*.

	Prentice-Hall, NJ.] attributes to Emshoff and Sisson (1970) [*"Design and Use of Computer Simulation Models". Macmillan, NY*], is based on the χ^2 goodness-of-fit test applied for selecting a time from which the numbers of observations below and above the running mean are equal in the statistical sense." "…seems to be quite simple and independent of any system-related parameter."
Positives	Simple. Independent of system related parameters [7].
Criticisms	None found in literature
Accuracy	Unknown
Simplicity	Good
Automation potential	Good
Generality (No assumptions)	Good
Parameters to estimate?	2 - Run length, batch size
Computation time	Good

Method ID	18
Method Name	Kelton and Law Regression Method
Brief description	Law & Kelton (1983) [11]: Make k independent replications of length m_0. Average over the replications to obtain the single time series $\overline{X}_1, \overline{X}_2, ..., \overline{X}_{m_0}$. Group the series into b batches and compute the b batch means. Using the Amemiya GLS procedure, fit a straight line through the batch means $\{p^*b + 1,…,b\}$, where p* is the maximum initial deletion proportion. Perform a test for zero slope at level β. If the test doesn't reject the null hypothesis of zero slope move the 'test window' backwards towards the start of the data series, redraw line and test for zero slope again. If the test rejects the null hypothesis at the first try, add data to create a longer run length and repeat procedure. If the test rejects the null hypothesis after the first try then truncation point has been found. (see Law & Kelton (1983) [11], p646-p647, for more detailed instructions)
Literature Review	Roth (1994) [6]:

	Systems Studied: 51 M/M/k queuing systems with ρ = 0.1 to 0.9 and k = 1 to max. Performance Criteria: Normal hypothesis test – H0: estimated mean = true mean (see paper [6] for details). Results: Kelton-Law method (18) failed in 43% of cases - probably because run length was not long enough!.. Pawlikowski (1990) [7]: "..can be applied only in the case of monotonic convergence to steady state" but "Keifer and Wolfowitz (1955) proved that in any stable, initially empty and idle GI/G/c queuing system, the mean delay in queue grows monotonically in time." "The procedure implementing this rule appears to be quite effective, especially in lowering the MSE of estimators (Roth 1985; Roth and Rutun 1985)." Kelton & Law (1983) [11]: Systems Studied: Parameter estimation carried out using an M/M/1 queue with ρ = 0.9, a triple tandem queue M/M/1/M/1/M/1, & a time shared computer model. Performance Criteria: (FBIAS) = Future bias, (PEBIAS) = point estimator bias, (COVER) & (EHL) = coverage and expected 1/2 length of CI for μ, (MAD) = mean absolute deviation of point estimator, replication length, execution time… Testing of procedure used 13 stochastic models with known μ: M/M/1 ρ = 0.8; 0.9; 0.95; M/M/1 LIFO ρ = 0.8; M/M/1 SIRO ρ = 0.8; M/M/1 L0=10 ρ = 0.8; E4/M/1 ρ = 0.8; M/H2/1 ρ = 0.8; M/M/2 ρ = 0.8; M/M/4 ρ = 0.8; M/M/1/M/1/M/1 ρ = (0.5, 0.7, 0.9); Time shared comp model; Central server comp model; Further tests carried out on 2 more "complex" problems. Results: Method performed quite well except for model M/M/1 ρ = 0.95. Average CI coverage = 84% out of a desired 90%. Usually reduced bias <= 1%. Law (1983) [34]: No tests done - just algorithm explained & commented on: "…shown to perform well for a wide variety of stochastic models. However, a theoretical limitation of the procedure is that it assumes that

	E[Y(i)] is a monotone function of i. This limits the overall applicability of this procedure". Linton & Harmonsky (2002) [36]: Model1: Queuing model with 2 servers in sequence. Inter-arrival time is exponential with mean of 8; server1 process time is exponential with a mean of 6; server2 process time is exponential with mean of 7. Model2: Same as for model1 except all distributions are triangular; inter-arrival time distribution is therefore defined as{min = 6, mode = 8, max = 10}; server1 process time defined as{min = 4, mode = 6, max = 8}; server2 process time as {min = 5, mode = 7, max = 9} Performance criteria: i) equality of variance (between methods) using Levene's test ii) equality of mean (between methods) using 2-sample t-test. Results: Found able to adjust to changes in distributions of the inter-arrival times and processing times. Computationally intensive. WARNING: Not tested to see if chosen truncation point was correct or efficient! Kimbler & Knight (1987) [46]: Systems Studied: M/M/1 queue with ρ = 0.9. This is a "highly congested system …has an unusually gradual transition into steady state" Performance Criteria: Steady state statistics (time in system) compared with theoretical solution using Dudewicz method for determining the best of k systems. Results: Complex to (code) execute. Method worked well. "…We also found that while the methods proposed by both Kelton (18) and Schruben (22) were very complex and sophisticated, the results which they produced did not warrant the additional level of difficulty" Gallagher et al (1996) [47]: Systems Studied: (based on Kelton and Law 1983) M/M/1 queues ρ = 0.8, 0.90, 0.95, 0.8(LIFO), 0.8 (L0 = 10), 0.8(Lq). E4/M/1 ρ = 0.8; M/M/2 ρ = 0.8; M/M/4 ρ = 0.8; Open model is 3 M/M/1 queues in tandem;

	Time-sharing computer; Central server computer. Outputs are waiting times; queue lengths; sum of waiting times; job response times; processing times. Performance Criteria: 1) Point estimator bias 2) mean abs deviation 3) realised coverage rates (replications method) 4) average CI half widths. Results: The MMAE (44) algorithm generally selected truncation points earlier in the output sequences than the Kelton and Law method (18). Law & Kelton (2000) [52]: "…a theoretical limitation of the procedure is that it basically makes the assumption that E(Yi) is a monotone function of i." Roth & Josephy (1993) [57]: Systems Studied: 30 M/Ek/1 and 24 Ek/M/1 queuing systems (which begin at rest); ρ varies from 0.25 to 0.925 by holding arrival rate fixed and varying service rate; k (parameter for the Erlang dist) varies from 1 to 20. Note: Experiments have fixed replication length. Performance Criteria: Bias = estimated mean - theoretical mean; CI size - a function of the variance; MSE; observed coverage probability of CIs. Results: "The relaxation time heuristic (17) satisfies the bias criterion in each experiment. The Kelton-Law (18) method, Cu-mean (3) and truncated mean (4) techniques are less consistent, causing rejection of the null hypothesis in 33, 10, and 3% of cases respectively." "The confidence interval coverage probabilities are quite consistent for each truncation rule." "K-L heuristic appears to be particularly successful in keeping the variance small…..suspect ..K-L heuristic yielded truncation pts smaller than those of the other rules." This is born out by the results!
Positives	The procedure implementing this rule appears to be quite effective, especially in lowering the MSE of estimators [7]. Found able to adjust to changes in distributions of the inter-arrival times and processing times. [36] Method worked well. [46] "K-L heuristic yielded truncation pts smaller than those of the relaxation time heuristic (17), Cu-mean (3) , and truncated mean (4)" [57]

Criticisms	Only fair performance for M/M/1 $\rho = 0.95$. [11] Assumes that E[Y(i)] is a monotone fn of i. This limits the overall applicability of this procedure" [52] [34] [7]. Computationally intensive.[36] Complex to (code) execute. [46]
Accuracy	Good/Medium
Simplicity	Poor
Automation potential	Good
Generality (No assumptions)	Medium – assumes monotonic behaviour – but "Keifer and Wolfowitz (1955) proved that in any stable, initially empty and idle GI/G/c queueing system, the mean delay in queue grows monotonically in time."[7]
Parameters to estimate?	9 – number of replications (k); initial run-length (m_0); number of points added to each replication if necessary (Δm); maximum replication length (m^*); number of batches (b); max initial deletion proportion (p^*); min initial deletion proportion (p_0); size of the test for zero slope (β); max number of segments over which a fit is made, including initial fit (f).
Computation time	Medium

Method ID	19
Method Name	Randomisation Tests for Initialisation Bias
Brief description	Mahajan & Ingalls(2004) [1]: Randomization tests are applied to test the null hypothesis that mean of the process is unchanged throughout the run. Null Hypothesis: No initialisation bias. 1. Let $Y_1, Y_2, \ldots, Y_m$ be a series of output from a run of length m. 2. Batch the data into b batches of length k 3. Obtain b batch means: $\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_b$ 4. Partition the batch means into 2 groups. For the 1st iteration the 1st group must include the 1st batch mean and the 2nd group contains the remaining b-1 batch means. 5. For each iteration compare the grand means of each group: Abs[Grand mean(grp1)-Grand mean(grp2)]. If the difference is significantly different from zero then the null hypothesis of no initial bias is rejected. To access the significance a distribution of difference is required. Since it is unknown, randomization is used to obtain an empirical distribution.

	6. If the hypothesis is rejected, the groups are rearranged so that the 2nd batch mean moves into the 1st group leaving b-2 batch means in the 2nd group. Step 5 is then repeated. 7. If the hypothesis is not rejected then the warm-up is in group1 and the steady state output is in group2.
Literature Review	Mahajan & Ingalls (2004) [1]: Systems Studied: Job Shop model consisting of 5 Cells, C(i), i=1…5. Each cell has different number of machines (resources). There are 3 customer classes A,B,C. Overall arrival rate is poisson. Service times are exponential with mean dependent on customer class and cells. Arriving parts are split into classes {A, B, C} with probability {0.5, 0.3, 0.2}. 3 types of utilisation are used: TypeI is high utilisation with an average utilisation of 90% and range 80-95%; TypeII is moderate with an average utilisation of 70% and range 65-80%; TypeIII is Low with average utilisation of 50% with range 45-65%. Models are started empty & idle. Initial run length is 1000 hrs, with 5 replications. Performance criteria: Final MSE & Variance, average computing time, percentage change in MSE, percentage change in variance. Method said to perform well if it reduces both MSE and Variance and is computationally efficient. Results: Recommended for use with highly utilised systems with a long run length. Also recommended for use with low utilised systems. "…no assumptions, like that of normality are required." Yucesan (1993) [23]: Generally more conservative than Conway Rule (10) or (graphical) Schruben (20). Performance of Randomisation tests (19) not satisfactory for case where output has high positive correlation. Advantage of this technique is that no assumptions are made regarding distribution of output.
Positives	Recommended for use with highly utilised systems with a long run length and with low utilised systems. [1] No assumptions are made regarding distribution of output. [1] [23]
Criticisms	Conservative. Performance not satisfactory when output has high positive correlation. [23] Fairly complicated and slow to run due to random shuffling.
Accuracy	Poor/Medium
Simplicity	Medium
Automation potential	Good
Generality (No assumptions)	Good

Parameters to estimate?	2 – run length (m); number of batches (b)
Computation time	Poor

Method ID	30
Method Name	Algorithm for a Static Dataset (ASD)
Brief description	Let $\tilde{F}_j(x \mid S(0))$ denote the empirical CDF of the k values from the independent replications. Let $x_{i,j}$ be the jth observation of the ith replication with $1 \le i \le k$ and $1 \le j \le n$. The sequence $\{x_{i,j}, i=1,\ldots,k\}$ can be considered as an independent random sample of X_i. The maximum difference of 2 CDFs X_1 and X_2 is given by $\max_x \lvert F_1(x)-F_2(x)\rvert$ where $F_i(x)$ is the proportion of X_i values less than or equal to x. 1. Calculate $\tilde{F}_j(x \mid S(0))$ for $1 \le j \le n$ by sorting $\{x_{i,j}, i=1,\ldots,k\}$. 2. Compute the max differences $\{d_j, j=1,\ldots,n-1\}$ of $\tilde{F}_j(x \mid S(0))$ and $\tilde{F}_n(x \mid S(0))$. 3. Compute for all j with $1 \le j \le n-1$ the number of differences which miss the threshold in the interval [j,n-1] 4. Choose truncation pt to be the minimum value of j after which only (a*100)% of the $d_j, d_{j+1}, \ldots, d_{n-1}$ exceed the threshold $z_{2,k;1-\alpha}$.
Literature Review	Bause & Eickhoff (2003) [4]: System Studied: Artificial processes: Linear transient mean; Linear transient variance; Exponential transient mean; ARMA(5,5); Periodic; Non-Ergodic. Results: Slower than methods Welch (6) & Cum-mean rule (3), but not significantly in practice; "ASD (30) and ADD (31) solve some problems of the methods of Fishman [Cum-mean rule (3)] and Welch (6) because they are based on the CDFs and not only the mean. Thus they take the definition of steady state better into account." "Even though the implementations of ASD and ADD (31) are not difficult, they are more complicated than the methods of Fishman (3) and Welch (6) . Their execution is more costly too. But this pays off when analyzing real-world models with complex transient behaviour." " Whenever steady state can not be deduced from a steady mean or the transient behaviour is not roughly known, ASD and ADD (31) are a more adequate choice. Their additional running time is acceptable since they are able to find a proper truncation point for a larger amount of models."

Positives	"..Based on the CDFs and not only the mean. Thus they take the definition of steady state better into account." ".. able to find a proper truncation point for a larger amount of models." [4]
Criticisms	Slower than methods Welch (6) & Cum-mean rule (3). [4]
Accuracy	Medium/Good (limited testing)
Simplicity	Medium
Automation potential	Good
Generality (No assumptions)	Good
Parameters to estimate?	4 - Replication number (k), run length (n), truncation criteria coefficient (a), Sig % α
Computation time	Medium

Method ID	31
Method Name	Algorithm for a Dynamic Dataset (ADD)
Brief description	1. Choose a ratio 1: r, and a level p ($0\leq p \leq 1$). 2. Initialise n to 0 3. Observe r+1 new X-intervals of all replications and compute the r+1 new random samples: $\{x_{i,n+1}, i=1,..,k\},\ldots,\{x_{i,n+r+1}, i=1,..,k\}$ 4. Set n = n+(r+1) 5. Set TS = $\{x_{i,n/(r+1)}, i=1,..,k\}$ 6. Compare TS with $\{x_{i,j}, i=1,..,k\}$ for j = [n/(r+1)]+1,…,n using the Kolmogoroff-Smirnov 2-sample test. 7. If more than (p×100)% of the compared random samples $\{x_{i,j}, i=1,..,k\}$ have a different probability distribution than TS: Goto 2. Otherwise terminate with truncation pt = n/(r+1)
	Bause & Eickhoff (2003) [4]: System Studied: Artificial processes: Linear transient mean; Linear transient variance; Exponential transient mean; ARMA(5,5); Periodic; Non-Ergodic. Results: Slower than methods Welch (6) & Cum-mean rule (3), but not significantly in practice; "ASD (30) and ADD (31) solve some problems of the methods of Fishman [Cum-mean rule (3)] and Welch (6) because they are based on the CDFs and not only the mean. Thus they take the definition of steady state better into

	account." "Even though the implementations of ASD and ADD (31) are not difficult, they are more complicated than the methods of Fishman (3) and Welch (6) . Their execution is more costly too. But this pays off when analyzing real-world models with complex transient behaviour." " Whenever steady state can not be deduced from a steady mean or the transient behaviour is not roughly known, ASD and ADD (31) are a more adequate choice. Their additional running time is acceptable since they are able to find a proper truncation point for a larger amount of models."
Positives	"..Based on the CDFs and not only the mean. Thus they take the definition of steady state better into account." ".. able to find a proper truncation point for a larger amount of models." [4] Analyses dynamically as output generated.
Criticisms	Slower than methods Welch (6) & Cum-mean rule (3). [4]
Accuracy	Medium/Good (limited testing)
Simplicity	Medium
Automation potential	Good
Generality (No assumptions)	Good
Parameters to estimate?	4 - Replication number (k), ratio parameter (r), truncation criteria coefficient (p), Sig % α
Computation time	Medium

Method ID	34
Method Name	Telephone Network Rule
Brief description	"Consider a network consisting of a set of nodes connected together by a series of links or paths, and suppose that these links represent cables through which telephone calls can travel from node to node within the system. At any given time, a series of telephone calls may be in progress within this network, each one associated with an origin node from which the call was initiated, and also with a destination node which ultimately receives the call. While the call is in progress it occupies a continuous pathway or route through this network connecting these origin and destination nodes. Every origin/destination node pair, (OD pair), is associated with a unique set of such connecting routes." "Our approach to determining the length of the start-up period involves comparing the observed arrival rate (rate at which calls are added to each OD pair) and departure rate (overall rate of departure from a given OD pair) of calls associated with each of the OD pairs in the network. If for each OD pair the difference

	between these arrival and departure rates is not statistically significantly different from 0 then we can be fairly confident that the system is close to steady state."
Literature Review	Zobel & Preston White Jnr (1999) [19]: Systems Studied: 4 representative 3-node networks. Arrival rates = (5,5,5/sec) & (8,3,1/sec), departure rates = 1 & 0.1/sec, link capacities = 4 & 20. Results: Unclear - paper does not seem to show how well method does!? But not a very general method as devised to be used specifically for telephone networks - though they briefly state that it may be useful in other types of queuing situations.
Positives	
Criticisms	Not a very general method as devised to be used specifically for telephone networks. Requires monitoring of each OD pair in model.
Accuracy	Unknown
Simplicity	Medium
Automation potential	Unknown
Generality (No assumptions)	Poor
Parameters to estimate?	None?
Computation time	Unknown

Method ID	37
Method Name	Glynn & Iglehart Bias Deletion Rule
Brief description	Let Y(t) = f(X(t)), where X = {X(t): $t \geq 0$} is a continuous-time Markov chain on state space S ={0,1,2…}. Let π_i be the stationary probabilities of X taken from the empirical steady-state distribution. Specifically let $\pi_i(t) = \frac{1}{t}\int_0^t I(X(s) = i)ds$ Let V(t) be the r.v with conditional distribution P{V(t)=i | X}= $\pi_i(t)$ And let T(t) = inf{s≥0:X(s)=V(t)} Delete from the data set {Y(s): 0≤s≤t}, all observations collected prior to T(t).

Literature Review	Glynn & Iglehart (1987) [38]: No testing of method carried out. Just explanation of method (theory). "[further work required], in order to determine the efficacy of the deletion estimator…"
Positives	
Criticisms	Very theoretical. Practical application not proven to be effective.
Accuracy	unknown
Simplicity	unknown
Automation potential	unknown
Generality (No assumptions)	unknown
Parameters to estimate?	unknown
Computation time	unknown

Method ID	38
Method Name	Wavelet-based spectral method (WASSP)
Brief description	Algorithm divides the initial output into a fixed number of batches of uniform size. Batch means are computed for all batches and a randomness test (von Neumann test) applied to the set of batch means. The randomness test is used to construct a set of spaced batch means such that the inter-batch spacer preceding each batch is sufficiently large to ensure all computed batch means are approx i.i.d so they can be tested for normality later. It is also used to determine an appropriate data truncation point – i.e. the inter-batch spacer preceding the first batch, beyond which all computed batch means are approximately independent of initial bias. (The rest of the method is about constructing CIs using wavelet based estimator)
Literature Review	Lada et al (2003) [39]: Systems Studied: Performance evaluated using a M/M/1 queue process with 90% server utilisation, exponential inter-arrival times with mean 10/9, exponential service times with mean 1 and empty-and-idle initial conditions. The steady state mean waiting time is 9. Results: Satisfied the various precision criteria set. This method, while deleting the warm-up, also produces CIs around an estimated mean value. Lada et al (2004) [54] & (2006) [56]: Further testing of the WASSP method:

	Systems Studied: (1) M/M/1 queue process with 90% server utilisation, exponential inter-arrival times with mean 10/9, exponential service times with mean 1 and empty-and-idle initial conditions. The steady state mean waiting time is 9. (2) AR(1) process with Normal errors. (3) AR(1)-to-Pareto (ARTOP) process. Results: "..test processes ..were specifically designed to explore the robustness of WASSP and its competitors against the statistical anomilies commonly encountered in the analysis of outputs generated from large scale steady-state simulation experiments." "..we believe WASSP represents an advance in spectral methods for simulation output analysis."
Positives	This method, while deleting the warm-up, also produces CIs around an estimated mean value.
Criticisms	
Accuracy	Unclear - as method is compared to batching methods rather than warm-up truncation methods
Simplicity	Medium
Automation potential	Good
Generality (No assumptions)	Good
Parameters to estimate?	Number of batches (k); batch size (m); (sig level α)
Computation time	Medium

Method ID	39
Method Name	Queueing approximations method (MSEASVT)
Brief description	Truncation point = $\underset{0 \le d < n}{\arg\min} \left\{ \hat{Bias}^2 \left[\hat{\vartheta} \right] + \hat{Var} \left[\hat{\vartheta} \right] \right\}$ Where d is the true truncation point, n is the total data length, $\hat{\vartheta}$ is the approximation of the true mean value wanted.
Literature Review	Rossetti & Delaney (1995) [41]: Systems Studied: A series of non-standard queuing models (i.e. not M/M/c) with traffic intensity ≥ 0.9. Ouput variable is waiting time. E2/E2/4; U/LN/3; initial conditions are either empty and idle, or

	stochastically set. Performance criteria: mean Absolute bias + 95% CIs. Results: The average performance of the method was superior to not truncating at all! “assumes no knowledge of the system based on pilot runs”
Positives	The average performance of the method was superior to not truncating at all! “assumes no knowledge of the system based on pilot runs” [41]
Criticisms	-
Accuracy	Unclear
Simplicity	Medium
Automation potential	Medium
Generality (No assumptions)	Medium - Applicable to steady state GI/G/m queueing models only.
Parameters to estimate?	3 – Mean ($\hat{\vartheta}$), variance (V), run length (n)
Computation time	Medium (could depend on how variance estimated)

Method ID	44
Method Name	Kalman Filter method (MMAE – Multiple Model Adaptive Estimation)
Brief description	Gallagher et al (1996) [47]: This method “estimates a state-space model for the simulation output, applies MMAE to provide a time-varying mean estimate, and selects a truncation point when the MMAE estimate is near a steady-state mean estimate.” (algorithm can be found in paper [47] and FORTRAN code is apparently available from the authors)
Literature Review	Gallagher et al (1996) [47]: Systems Studied: (based on Kelton and Law 1983) M/M/1 queues ρ = 0.8, 0.90, 0.95, 0.8(LIFO), 0.8 (L0 = 10), 0.8(Lq). E4/M/1 ρ = 0.8; M/M/2 ρ = 0.8; M/M/4 ρ = 0.8; Open model is 3 M/M/1 queues in tandem; Time-sharing computer; Central server computer.

	Outputs are waiting times; queue lengths; sum of waiting times; job response times; processing times. Performance Criteria: 1) Point estimator bias 2) mean abs deviation 3) realised coverage rates (replications method) 4) average CI half widths. Results: Failure rate for the MMAE (44) algorithm was small apart from for the M/M/1 ρ = 0.95 model and M/M/4 model. Point estimate bias is small. Mean absolute deviation values are relatively small except for the M/M/1 model with ρ = 0.95. Coverage rates are near nominal rate of 0.9 except for the M/M/1 with ρ = 0.95. Average half widths are relatively small. This new algorithm "…appears to result in truncated sequences with better estimation characteristics than sequences truncated by existing methods". " The MMAE (44) algorithm generally selected truncation points earlier in the output sequences than the [Kelton and Law method (18)]. …The MMAE (44) truncated sequences resulted in mean estimates with slightly larger bias and smaller mean average deviation, higher coverage rates and smaller average half widths (as there were more observations to use than for Kelton and Law method (18) truncated sequences)" Law & Kelton (2000) [52]: In reviewing Gallagher, Bauer and Maybeck's paper [47] this book comments that "…they tested on 12 queuing models. For the delay-in-queue process for the M/M/1 queue with ρ = 0.8 and 0.9, their procedure produced average warm-up periods of L = 180 and 406 when the run length was equal to 1500. In the latter case, E(Di) differs from the steady-state mean by less than 1 percent for i ≥ 780 (approx)…" (So procedure under estimates warmup? - not explained in book).
Positives	"...appears to result in truncated sequences with better estimation characteristics than sequences truncated by existing methods". "The MMAE algorithm generally selected truncation points earlier in the output sequences [than the Kelton and Law method (18) method]" [47]
Criticisms	Procedure underestimates warmup period for model M/M/1 rho=0.9. [52] Did not perform well for queue models with high rho e.g. 0.95. [47]
Accuracy	Medium
Simplicity	Poor
Automation potential	Medium
Generality (No assumptions)	Good

Parameters to estimate?	7 – Variance (V), $\bar{\bar{y}}, \hat{\phi}_1, \hat{\phi}_2, k^*$ (see Gallagher et al (1996)[47]) and run length(n) and Number of replications (M).
Computation time	Medium/Poor ?

(Return to top of document)

GRAPHICAL METHODS

Method ID	1
Method Name	Simple Time Series Inspection
Brief description	Draw graph of time series and truncate where line appears to "flatten out".
Literature Review	Gordon (1969) [48] advises that the cumulative mean (i.e. replications – see method Cumulative-Mean Rule (3)) be used in preference to this simple method because of the large variation between replication outputs.
Positives	Very simple and quick. Applicable to all systems. No assumptions or parameters to estimate.
Criticisms	Subjective. Therefore accuracy probably affected by experience of user. Large variation between single replication runs causes different truncations points for different runs [48]. Large variation in data output makes it difficult to judge where "flatness" occurs. Requires user intervention/judgement. Automation would require invention of automatic method to judge "flatness" of line.
Accuracy	Poor
Simplicity	Good
Automation potential	Poor
Generality (No assumptions)	Good
Parameters to estimate?	None
Computation time	Good

Method ID	2
Method Name	Ensemble (Batch) Average Plots
Brief description	Perform replications. Calculate batch means within each replication. Average corresponding batch means across replications and plot them. Choose as the truncation point the point where observations appear to vary around a common mean.
Literature Review	Algorithm explained in Banks et al (2001) [51]
Positives	Very simple and quick. Applicable to all systems. No assumptions. Batching and replications may reduce autocorrelation.
Criticisms	Subjective. Therefore accuracy probably affected by experience of user. The parameter, batch size, needs to be estimated. The batch size can affect accuracy. Requires user intervention/judgement. Automation would require invention of automatic method to judge "flatness" of line.
Accuracy	Poor
Simplicity	Good
Automation potential	Poor
Generality (No assumptions)	Good
Parameters to estimate?	2 – batch size, number of replications
Computation time	Good

Method ID	3
Method Name	Cumulative-Mean Rule
Brief description	Perform replications. Take the mean of each replication. Plot cumulative mean w.r.t. number of observations (n). Choose truncation point as point where cumulative mean becomes stable.
Literature Review	Bause & Eickhoff (2003) [4]: "The method of Fishman [i.e. Cumulative-Mean Rule (3)], smoothes the simulation data by calculating the mean at each time index. This method is easy to implement and therefore very popular. It creates expressive plots for simple transient behaviour. But analysis of the

	steady-state phase just on the basis of the mean might lead to problems. As Welch remarked (Welch 1983 "The statistical analysis of simulation results", The computer performance modelling handbook. Ed. S Lavenberg, Academic Press 268-328) convergence of the mean is a necessary but not a sufficient condition for stationarity. Therefore this method is not suitable for the analysis of complex transient behaviour. In such cases it is advisable to compare the results with a plot of the original data." "The method of (Fishman) Cumulative-Mean Rule (3) is very useful when estimating a steady mean. If the transient behaviour is roughly known this method is adequate". Roth (1994) [6]: Systems Studied: 51 M/M/k queuing systems with ρ = 0.1 to 0.9 and k = 1 to max. Performance Criteria: Normal hypothesis test – H0: estimated mean = true mean (see paper [6] for details). Results: This method, Cumulative-Mean Rule (3), failed the bias criteria just twice (out of 400 experiments, each comprising of 40 replications) for ρ = 0.9. Nelson (1982) [16] explains this algorithm & comments that it is "subjective". Gafarian et al. (1978) [32]: Systems Studied: M/M/1 queue with output variable as waiting time in queue per customer; ρ = 0.1, 0.5, 0.7, 0.9. Method is tested 100 times. 'Goodness' (Performance) Criteria: (xvi) Accuracy: The ratio, mean of [estimated truncation values] / [true truncation value], is calculated and a value near 1 implies method accurate. (xvii) Precision: Measure of variation in estimated truncation value is calculated as Sqrt(variance of estimated truncation points) / (mean of estimated truncation points). A value close to zero implies method precise. (xviii) Generality: Judged to be general if the rule performs well across a broad range of systems and parameters within a system. (xix) Cost: Expense in computer time. (xx) Simplicity: Accessible to average practitioner.

	(i) Accuaracy, (ii) Precision and (iii) Generality are considered first. Any method not satisfactory on all three is discarded. Computer cost is a last priority. WARNING: Only ONE test run carried out! Results: Simplicity criteria satisfied. Very poor results. Fails accuracy criteria: Grossly overestimates truncation point - large positive bias. Wilson & Pritsker (1978) [35] comments that this rule requires user judgment. "Law (1975)…found that" this method "grossly overestimated the number of observations to be deleted - "not surprising" as it uses cumulative statistics which react slowly to changes in system status. However Law based his results on a single application of the rule rather than considering random variation in truncation point. Gafarian et al [32] applied this method to an M/M/1/infinity queue, ρ = 0.1, 0.5, 0.7, 0.9. Found it yielded large truncation pts and took excessive computing time. Fishman (2001) [37] describes the cumulative-mean algorithm. Sandikci & Sabuncuoglu (2006) [45]: Systems Studied: Uses 2 types of manufacturing system model: (1) serial production lines (2) job-shops. Output variable is 'time in system'. Results: Subjective and can tend to over estimate truncation point. Gordon (1969) [48] describes the cumulative-mean rule algorithm and comments that this method is preferable to the simple method (1) because of the large variation between replication outputs. Banks et al (2001) [51] briefly describes cumulative-mean rule and criticises cumulative average methods in general: "…become less variable as more data are averaged. Therefore it is expected that the left side of the curve will always be less smooth than the right side. More importantly, cumulative averages tend to converge more slowly to long-run performance than do ensemble averages, because cumulative averages contain all observations, including the most biased ones form the beginning of the run. For this reason cumulative averages should be used only if it is not feasible to compute ensemble averages such as when only a single replication is possible".

	Roth & Josephy (1993) [57]: Systems Studied: 30 M/Ek/1 and 24 Ek/M/1 queuing systems (which begin at rest); ρ varies from 0.25 to 0.925 by holding arrival rate fixed and varying service rate; k (parameter for the Erlang dist) varies from 1 to 20. Note: Experiments have fixed replication length. Performance Criteria: Bias = estimated mean - theoretical mean; CI size - a function of the variance; MSE; observed coverage probability of CIs. Results: "The Kelton-Law method (18), Cum-mean (3) and truncated mean (4) techniques are less consistent [than the relaxation time heuristic (17)], causing rejection of the null hypothesis in 33, 10, and 3% of cases respectively." "The confidence interval coverage probabilities are quite consistent for each truncation rule."
Positives	Easy to implement and therefore very popular. It creates expressive plots for simple transient behaviour [4].
Criticisms	Subjective [16][45]. Therefore, accuracy probably affected by experience of user. Uses cumulative statistics which react slowly to changes in system status [35][51]. Cumulative averages tend to converge more slowly to long-run performance than do ensemble averages [35]. Hence tends to overestimate truncation point with a large positive bias [32] [35] [45]. Requires user intervention/judgement. Automation would require invention of automatic method to judge "flatness" of line.
Accuracy	Poor
Simplicity	Good
Automation potential	Poor
Generality (No assumptions)	Good
Parameters to estimate?	1 – number of replications
Computation time	Good

Method ID	4
Method Name	Deleting-The-Cumulative-Mean Rule
Brief description	Perform replications. Calculate batch mean within each replication. Average corresponding batch means

	across replications. Plot cumulative mean w.r.t. number of batches (n). Delete initial batch means one by one, calculating new cumulative mean until initialisation bias is removed. Can do this by eye or using a criteria for 'closeness' of the batch means.
Literature Review	Roth & Josephy (1993) [57]: Systems Studied: 30 M/Ek/1 and 24 Ek/M/1 queuing systems (which begin at rest); ρ varies from 0.25 to 0.925 by holding arrival rate fixed and varying service rate; k (parameter for the Erlang dist) varies from 1 to 20. Note: Experiments have fixed replication length. Performance Criteria: Bias = estimated mean - theoretical mean; CI size - a function of the variance; MSE; observed coverage probability of CIs. Results: "The Kelton-Law method (18), Cum-mean (3) and truncated mean (4) techniques are less consistent [than the relaxation time heuristic (17)], causing rejection of the null hypothesis in 33, 10, and 3% of cases respectively." "The confidence interval coverage probabilities are quite consistent for each truncation rule."
Positives	Fairly simple. Uses the truncated sample mean as the testing instrument in order to correct for the slow reaction of cumulative means.
Criticisms	Several parameters required. Preliminary run required.
Accuracy	Poor/Medium
Simplicity	Good
Automation potential	Medium
Generality (No assumptions)	Good
Parameters to estimate?	4 – batch size, replication number, closeness parameter, and run length or batch number.
Computation time	Good

Method ID	5
Method Name	CUSUM Plots
Brief description	Calculate and plot the Cumulative Sum Statistic, S, for one long run divided into batches (or can use batches and replications).

	$S_j = \sum_{i=1}^{j} (\bar{Y} - Y_i)$, where j = 1,..,m, where m is the total number of batches. S_0=S_m=0. With no initial bias the values for S_j, j not equal to 0 or m, should appear to vary around zero. With initial bias present S_j values are expected to be consistently above or below zero.
Literature Review	Nelson (1982) [16] "Schruben derived a plot that can be formed from a single long replication, and that is particulary sensitive to the presence of initial-condition bias." www.itl.nist.gov: In reference to CUSUM charts: while not as intuitive and simple to operate as Shewhart charts, CUSUM charts have been shown to be more efficient in detecting small shifts in the mean of a process. In particular, analyzing Average Run Length's for CUSUM control charts shows that they are better than Shewhart control charts when it is desired to detect shifts in the mean that are 2 sigma or less.
Positives	Simple. A sensitive test.
Criticisms	Preliminary run required.
Accuracy	Unknown
Simplicity	Good
Automation potential	Poor
Generality (No assumptions)	Good
Parameters to estimate?	Batch size
Computation time	Good

Method ID	6
Method Name	Welch's Method
Brief description	Produce k replications of run length n. Calculate the n means across these k replications. For a given time window w, plot the moving averages against number of observations (1,..,n). If the plot is reasonably smooth (judged by eye), choose as the truncation point the point on the graph where line becomes 'flat' (judged by eye). If not smooth enough, choose a different window (w) size and redraw plot etc...
Literature Review	Mahajan & Ingalls (2004) [1]:

	<u>Systems Studied</u>: Job Shop model consisting of 5 Cells, C(i), i=1…5. Each cell has different number of machines (resources). There are 3 customer classes A,B,C. Overall arrival rate is poisson. Service times are exponential with mean dependent on customer class and cells. Arriving parts are split into classes {A, B, C} with probability {0.5, 0.3, 0.2}. 3 types of utilisation are used: TypeI is high utilisation with an average utilisation of 90% and range 80-95%; TypeII is moderate with an average utilisation of 70% and range 65-80%; TypeIII is Low with average utilisation of 50% with range 45-65%. Models are started empty & idle. Initial run length is 1000 hrs, with 5 replications. <u>Performance criteria</u>: Final MSE & Variance, average computing time, percentage change in MSE, percentage change in variance. Method said to perform well if it reduces both MSE and Variance and is computationally efficient. <u>Results</u>: Method didn't work well for most experiments - It is highly subjective. Bause & Eickhoff (2003) [4]: <u>System Studied</u>: Artificial processes: Linear transient mean; Linear transient variance; Exponential transient mean; ARMA(5,5); Periodic; Non-Ergodic. <u>Results</u>: "The method of Welch (6) is based on the column average… It has the same advantages as the method of Fishman [i.e. Cum-mean rule (3)], but suffers also from the same disadvantages." "This method is easy to implement and therefore very popular. It creates expressive plots for simple transient behaviour. But analysis of the steady-state phase just on the basis of the mean might lead to problems. As Welch remarked (Welch 1983 "The statistical analysis of simulation results", The computer performance modelling handbook. Ed. S Lavenberg, Academic Press 268-328) convergence of the mean is a necessary but not a sufficient condition for stationarity. Therefore this method is not suitable for the analysis of complex transient behaviour. In such cases it is advisable to compare the results with a plot of the original data." "The extension is that the column averages are smoothed again by calculating the means of the sliding window. This gives a better distinction between the random and the systematic error. But smoothing might lead to inaccurate results: the moving average is calculated from the means at different points in time and in general the process changes over time. E.g. smoothness of the kink depends on window size. A more serious problem occurs e.g. when analysing periodic processes especially if the window size conforms to the cycle length." "The methods of Fishman (3) and Welch (6) are very useful when estimating a steady mean. If the transient behaviour is roughly known these two methods are adequate".

	Pawlikowski (1990) [7]: "…as has been indicated by Conway (1963), accumulative statistics such as running mean usually stabilize very slowly with time, and therefore usually give over estimated values of.." the warm-up length. Law (1983) [34]: No tests done (just an illustration of method) - algorithm explained & commented on: "The parameters need to be determined by trial and error…..One drawback of Welch's procedure is that it might require a large number of replications to make the plot reasonably stable if the process is highly variable." Linton & Harmonsky (2002) [36]: Model1: Queuing model with 2 servers in sequence. Inter-arrival time is exponential with mean of 8; server1 process time is exponential with a mean of 6; server2 process time is exponential with mean of 7. Model2: Same as for model1 except all distributions are triangular; inter-arrival time distribution is therefore defined as{min = 6, mode = 8, max = 10}; server1 process time defined as{min = 4, mode = 6, max = 8}; server2 process time as {min = 5, mode = 7, max = 9} Performance criteria: i) equality of variance (between methods) using Levene's test ii) equality of mean (between methods) using 2-sample t-test. Results: Produced comparable results with relaxation heuristic method (17). Found able to adjust to changes in distributions of the inter-arrival times and processing times. More practical method especially as not based on any assumptions about type of system. WARNING: Not tested to see if chosen truncation point was correct or efficient! Sandikci & Sabuncuoglu (2006) [45]: Systems Studied: Uses 2 types of manufacturing system model: (1) serial production lines (2) job-shops. Output variable is 'time in system'. Results: "Instead of cumulative averages plot, one can think of using Welch's technique due to its popularity. However (Fig1 [see [45]] shows that) these two techniques do not produce significantly different results. Besides, Welch's technique requires the analyst to decide on a window size (w) by trial and error which makes it practically less applicable."

	Banks et al (2001) [51]: Briefly touches upon Welch method as one of several smoothing methods. Law & Kelton (2000) [52]: Sets out the Welch method with instructions and recommendations on choosing the parameters. "The major difficulty in applying Welch's procedure in practice is that the required number of replications may be relatively large if the process is highly variable. Also the choice of [truncation point] is somewhat subjective". Alexopoulos & Seila (1998) [53]: Sets out the Welch method with instructions and an example. "It should be noted that the method of Welch may be difficult to apply in congested systems with output time series having autocorrelation functions with very long tails."
Positives	Easy to implement and therefore very popular. It creates expressive plots for simple transient behaviour [4]. Found able to adjust to changes in distributions of the inter-arrival times and processing times. More practical method especially as not based on any assumptions about type of system [36].
Criticisms	Highly subjective [1] [52]. Smoothing might lead to inaccurate results as the moving average is calculated from the means at different points in time and in general the process changes over time. This can particularly be a problem when analysing periodic processes especially if the window size conforms to the cycle length [4]. Accumulative statistics such as running mean usually stabilize very slowly with time, and therefore usually give a conservative estimate of warm-up period. Might require a large number of replications to make the plot reasonably stable if the process is highly variable [34] [52]. Cumulative averages plot, and Welch do not produce significantly different results. Welch's technique requires the analyst to decide on a window size (w) by trial and error which makes it practically less applicable [45]. The parameters need to be determined by trial and error [34]. The method of Welch may be difficult to apply in congested systems with output time series having autocorrelation functions with very long tails [53]. Requires automation of judging when line is ‘smooth’ (i.e. window size estimation) and judging when line is ‘flat’.
Accuracy	Poor
Simplicity	Good
Automation potential	Poor
Generality (No assumptions)	Good

Parameters to estimate?	3 - smoothing average window size (w), number of replications (k) and run length (n)
Computation time	Good

Method ID	7
Method Name	Variance Plots (or Gordon's Rule)
Brief description	Gordon (1969) [48] "In the absence of initial bias the standard deviation can be expected to be inversely proportional to square root of n", where n is the number of observations. "By examining the way the standard deviation changes with sample length, it is possible to see whether this relationship is being met." Make k replications to estimate the variance. Plot log(standard deviation) against log(n). Take as the truncation point the point where the graph "approximates a straight line sloping downwards at the rate of 1 in 2 (for equi-scaled axes)."
Literature Review	Pawlikowski (1990) [7]: "This rule was analysed in Gafarian et al 1978 and Wilson & Pritsker 1978b using [a specific variance estimator (see these papers [32][35])]….In this case the rule can give an overestimated value of [warm-up length]. No results have been published on the effectiveness of this rule when more accurate estimators of [variance] are applied." Gafarian et al. (1978) [32]: Systems Studied: M/M/1 queue with output variable as waiting time in queue per customer; ρ = 0.1, 0.5, 0.7, 0.9. Method is tested 100 times. 'Goodness' (Performance) Criteria: (i) Accuracy: The ratio, mean of [estimated truncation values] / [true truncation value], is calculated and a value near 1 implies method accurate. (ii) Precision: Measure of variation in estimated truncation value is calculated as Sqrt(variance of estimated truncation points) / (mean of estimated truncation points). A value close to zero implies method precise. (iii) Generality: Judged to be general if the rule performs well across a broad range of systems and parameters within a system. (iv) Cost: Expense in computer time. (v) Simplicity: Accessible to average practitioner.

	(i) Accuaracy, (ii) Precision and (iii) Generality are considered first. Any method not satisfactory on all three is discarded. Computer cost is a last priority. WARNING: Only ONE test run carried out! Results: Simplicity criteria satisfied. Fails accuracy criteria: Grossly overestimates truncation pt - large positive bias. Wilson&Pritsker (1978) [35]: Requires user judgment. Law 1975 found that this method grossly overestimated deletion point - "not surprising" as it uses cumulative statistics which react slowly to changes in system status. However Law based has results on a single application of the rule rather than considering random variation in truncation point. Gafarian et al [32] applied this method to an M/M/1/infinity queue. ρ = 0.1, 0.5, 0.7, 0.9. Found it yielded large truncation pts and took excessive computation time. Gordon (1969) [48]: Instructional book - Describes variance plots algorithm. "The chosen values are probably conservative because they are accumulated statistics."
Positives	Simple [32]
Criticisms	Rule can give an overestimated value of warm-up length. (No results have been published on the effectiveness of this rule when more accurate estimators of variance are applied) [7]. Grossly overestimates truncation pt - large positive bias [32]. Requires user judgment. Law 1975 found that this method grossly overestimated deletion point - "not surprising" as it uses cumulative statistics which react slowly to changes in system status. However Law based has results on a single application of the rule rather than considering random variation in truncation point [35].
Accuracy	Poor
Simplicity	Good
Automation potential	Medium
Generality (No assumptions)	Good
Parameters to estimate?	2 - number of replications (k) and variance (V)
Computation time	Good

Method ID	8

Method Name	Statistical Process Control Method (SPC)
Brief description	Robinson (2005) [8]: Perform k replications of run-length n. Collect output data and calculate n means over the k replications. Test that this output meets the assumptions of SPC – that the data are normally distributed and not correlated – by calculating batch means (est. batch length m using Von Neumann (1941) test for correlation); then using a normality test (e.g. Anderson-Darling (1954)). If number of batches (b) falls below 20 carry out further model runs. Once assumptions met, construct a control chart for the batched data. Estimate the population mean and standard deviation from the 2nd half of the time series (assumption: that mean and standard deviation stable in 2nd half of data). Identify the initial transient by viewing the chart and identifying the point at which the time series data are in control and remain in control.
Literature Review	Mahajan & Ingalls (2004) [1]: Systems Studied: Job Shop model consisting of 5 Cells, C(i), i=1…5. Each cell has different number of machines (resources). There are 3 customer classes A,B,C. Overall arrival rate is poisson. Service times are exponential with mean dependent on customer class and cells. Arriving parts are split into classes {A, B, C} with probability {0.5, 0.3, 0.2}. 3 types of utilisation are used: TypeI is high utilisation with an average utilisation of 90% and range 80-95%; TypeII is moderate with an average utilisation of 70% and range 65-80%; TypeIII is Low with average utilisation of 50% with range 45-65%. Models are started empty & idle. Initial run length is 1000 hrs, with 5 replications. Performance criteria: Final MSE & Variance, average computing time, percentage change in MSE, percentage change in variance. Method said to perform well if it reduces both MSE and Variance and is computationally efficient. Results: Recommended for use with highly utilised systems with a long run length. Robinson (2005) [8]: Systems Studied: AR(1) with Normal errors and Exponential errors; M/M/1 with arrival rate = 0.8, service rate=1, starts empty & idle; GSS M/M/1. Results: Performs well. Slight concern for small sample sizes. Simple to implement but not easily automatable; assumes normality & low autocorrelation (though some what alleviated by batching & reps); need to estimate 5 parameters.

	Law & Kelton (2000) [52]: In reviewing Robinson's paper [8] this book comments that "…He tested the procedure on several stochastic models including the delay-in-queue process D1,D2,…for the M/M/1 queue with $\rho = 0.8$. The procedure produced average warm-up periods of L = 502 and 1006 when the run length was equal to 2000 and 4000 respectively. Since E(Di) differs from the steady-state mean by less than 1 percent for $i \geq 182$, it would appear that the SPC procedure produces conservative estimates of the warm-up period for this particular problem."
Positives	Simple [8]. Recommended for use with highly utilised systems with a long run length [1].
Criticisms	Not easily automatable; assumes normality & low autocorrelation (though some what alleviated by batching & replications); need to estimate 5 parameters. Slight concern over performance for small sample sizes [8]. Appears that the SPC procedure produces conservative estimates of the warm-up period for delay-in-queue process D1,D2,…for the M/M/1 queue with $\rho = 0.8$. [52]
Accuracy	Medium
Simplicity	Good
Automation potential	Poor
Generality (No assumptions)	Medium - assumes normality & low autocorrelation (though both some what alleviated by batching & reps)
Parameters to estimate?	5 – number of replications (k); run-length (n); batch size (m); mean and variance (V) of data.
Computation time	Good

(Return to top of document)

HYBRID TESTS

Method ID	27
Method Name	Pawlikowski's Sequential Method
Brief description	Based on the Optimal test (22).

	First use any graphical or heuristic method to estimate the truncation point and delete data up to this point. Then, test the stationarity of the next n_t observations (collecting more observations if necessary to do so) using the Optimal test (22). If the sequence is not deemed stationary, delete a number of the observations from the beginning of the tested sequence, collect more observations if necessary and retest another n_t observations from the beginning of the sequence. Carry on till stationarity is found.
Literature Review	Pawlikowski (1990) [7]: No testing of method. Paper details creation of a Hybrid of a graphical method/heuristic method and the Optimal test (22). Computer code is written down in this paper. "Despite the sophisticated theory hidden behind these tests, they appear to be simple numerically and can be applied to a wide class of simulated processes"…"The main practical problem with their implementation is that they require a priori knowledge of the variance estimator of the simulated process in steady state."
Positives	"..appear to be simple numerically and can be applied to a wide class of simulated processes" [7]
Criticisms	"..require a priori knowledge of the variance estimator of the simulated process in steady state." [7]
Accuracy	Depends on first bias test used. If graphical or heuristic method overestimated truncation point in first place then a conservative warm-up period would be accepted without considering a shorter period.
Simplicity	Medium - depends on first bias test used.
Automation potential	Depends on first bias test used. And need to either decide if bias is positive or negative or use the two-sided test for optimal test.
Generality (No assumptions)	Medium – Asymptotic assumptions a possible problem – need to either decide if bias is positive or negative or use the two-sided test. And depends on first bias test used.
Parameters to estimate?	? – from first graphical or heuristic truncation point method 4/5 – from Optimal test[22]: Variance (V), degrees of freedom (d), run length (n), number of reps (k). May need to use batches (b) 7 - from hybrid algorithm: n_{max} the max allowed length of run; $n_{0,max}$ the max allowed length of the warm-up period; n_v the length of sequence used for estimating the steady-state variance; n_t the length of the sequence tested for stationarity; α_t sig level; γ_v the safety coefficient for the estimate of variance to represent the steady state; γ the exchange coefficient determining the number of new observations included in each sequential test for stationarity (i.e. how many obs to discard each time)
Computation time	Depends on initialisation bias methods used

Method ID	28
Method Name	Scale Invariant Truncation Point Method (SIT)
Brief description	Let $X_1, X_2, \ldots, X_n$ be a set of observations. Divide this set into b (b>3) equal batches of m points. Start with a small value for $m = m_{init}$ and increase m by a multiplicative factor Δm until a test for bias shows {true, false, false} when applied to the first three batches. "More complicated and intelligent search strategies may give improved results". The tests for bias could be any 'good' or combination of initialisation bias tests e.g. the Optimal test (22).
Literature Review	Jackway & deSilva (1992) [17]: Systems Studied: Model1: M/M/1 queue system with $\rho = 0.9$. The output variable is 'waiting time'. Model2: 3 capacitated M/M/s queues with feedback. The output variable is 'number of customers in system'. Model3: Time-share computer system. The output variable is response time. Model4: Central server computer system. The output variable is response time. Model5: AR(2) time series with artificial bias: $X(i) = 0.75X(i-1)-0.5X(i-2)+e(i)+b(i)$, where $e(i)\sim N(0,1)$ & $b(i) = (i-30)/30, i \leq 30$. Results: Results not very convincing. The average truncation pt for SIT method is compared with 3 other methods' average truncation pt results as well as the theoretical value. All methods majorly underestimate with SIT being the best of the methods (but not enough info given to identify other methods at this time AND these methods may not have been optimising the same criteria therefore irrelevant comparison?). Mean percentage of runs (out of 10) where bias was detected given for SIT for each of the 5 models tested: range from 40% for model3 to 100% for models 2 & 5.
Positives	-
Criticisms	Variable results.
Accuracy	Unknown - Not well tested.
Simplicity	Good/Medium

Automation potential	Depends on procedures used
Generality (No assumptions)	Depends on initialisation bias methods used
Parameters to estimate?	3 - Run length (n), Δm – multiplicative factor, m_{init} – starting size of batches Plus all those from other procedures used.
Computation time	Depends on initialisation bias methods used