

# NAEEEC National Appliance & Equipment Checktesting Program

## Statistical Basis for the Determination of Checktesting Validity Criteria

Report prepared by:  
Professor Robert Bartels (University of Sydney)  
and Lloyd Harrington, Energy Efficient Strategies  
for NAEEEC

Original Report: January 1999  
Updated and corrected: February 2004

This document has been prepared for the appliance industry and energy efficiency regulators in Australia to assist in the interpretation of checktesting results. It provides a theoretical background to the verification process and provides the basis for checktesting rules that have been adopted by NAEEEC.

Full details of the checktesting process are described in the Administrative Guidelines which are available from [www.energyrating.gov.au](http://www.energyrating.gov.au). The Administrative Guidelines also set out the verification limits and applicable tolerances that are applied during checktests within the energy labelling and MEPS program. A copy of this paper is also available from this website.

### 1. Background

The energy labelling program for major household electrical appliances has been operational in various states across Australia since 1986. Currently the national scheme covers dishwashers, clothes washers, clothes dryers, refrigerators, freezers and single phase non-ducted air conditioning units. It is mandatory for appliances covered by the scheme to carry an approved label before they can be offered or displayed for retail sale.

Similarly, Minimum Energy Performance Standards (MEPS) were introduced nationally for refrigerators, freezer and main pressure electric storage water heaters in October 1999. Since that time MEPS has been progressively introduced for a range of residential, commercial and industrial equipment including single and three phase air conditioners, three phase electric motors, ballasts for fluorescent lamps, linear fluorescent lamps, distribution transformers and commercial refrigeration. In addition, MEPS levels for existing products are subject to periodical review and levels are revised from time to time (e.g. refrigerators and freezers in 2005, small electric water heaters in 2005, electric motors in 2006, air conditioners in 2007).

The latest information on products that are within the energy labelling and MEPS programs can be found on [www.energyrating.gov.au](http://www.energyrating.gov.au)

Registration for energy labelling and MEPS is mandatory. To obtain registration of a product, manufacturers are generally required to submit test reports<sup>1</sup> to demonstrate compliance with the requirements of the relevant Australian Standard. The veracity of the energy consumption, efficiency and performance values claimed in these reports are usually accepted on initial application without requirement for verification through independent testing.

An essential output of the scheme is ensuring that manufacturers' energy efficiency and appliance performance claims accurately reflect the information contained within their original application for registration. This verification process is known as checktesting and is effectively the major quality assurance procedure for the energy labelling and MEPS schemes.

Consumers in Australia regard the energy labelling and MEPS programs as highly credible and it is essential that this level of confidence is maintained. Checktesting is fundamental to protecting the interests of the various stakeholders and maintaining the credibility of the programs.

### **1.1 Purpose of this Paper**

The purpose of this report is to define validity criteria, based on a sound statistical approach, for determining whether a check tested unit complies with energy or performance measures declared by the manufacture or whether it meets the minimum performance requirements set out in the relevant Australian standard.

### **1.2 Definitions used in this report**

For the purpose of this report, the following definitions apply:

**Application for Registration:** An application by a manufacturer or importer submitted to a state regulator for energy labelling or MEPS. The requirements are defined in the relevant Australian Standard for the particular product as well as state regulations. An application for registration includes the details of the applicant, the product, test reports (as applicable) and other relevant data. The technical and performance data which must be supplied with the application are defined in the regulatory standard for each product. When a manufacturer submits an application for registration, they are effectively declaring the energy and performance characteristics of the product. They are also required to declare that they meet the requirements stipulated in the relevant Australian Standard.

**Manufacturer Declaration:** A declaration of energy or performance made either within an application for registration or through manufacturer information supplied with the product (accompanying literature, user manuals, affixed to the product) or at the point of sale (including advertising).

---

<sup>1</sup> The requirements regarding the number of products that have to be tested and the details to be submitted for registration vary by product. The relevant standard sets out the detailed requirements.

**Performance Measure:** description of the energy service provided by the product in terms of capacity or ability to perform a specified function (eg dry clothes, wash dishes, keep food cool, light output, shaft power etc.).

**Performance Limit:** The minimum required level of a performance measure specified in the relevant Australian Standard. Generally, performance limits are determined by the relevant standards committee and are set at a level that provides a level of energy service that could reasonably be expected by a user. Performance limits are set primarily for consumer protection. Declarations of energy consumption (or energy efficiency) have little meaning without a specification of the performance measure (energy service) associated with the energy consumption.

**Minimum Energy Performance Standard (MEPS):** A performance limit that relates specifically to the energy consumption or energy efficiency of a product.

**Repeatability:** The ability to replicate a test result (i.e. an energy or other performance measure) using the same materials, personnel and test equipment on the same product in the same test laboratory.

**Reproducibility:** The ability to replicate a test result (i.e. an energy or other performance measure) using the different materials, personnel and test equipment on the same product in a different test laboratory.

**Standard Deviation:** Values of standard deviation quoted for actual product measurements in this report generally refer to the sample standard deviation which is calculated as follows:

$$\text{Sample Standard Deviation} = \sqrt{\frac{n \sum x^2 - (\sum x)^2}{n(n-1)}}$$

**Verification:** Note that there are two types of verification that occur during a check test:

- (i) Verification of a manufacturer's declaration (e.g. energy, volume, capacity etc.).
- (ii) Verification that a performance limit specified in the standard is achieved by the relevant model (includes MEPS in the case of energy).

## 2. Verification Issues

### 2.1 Factors that will affect the verification process

Very little information is generally available on the different factors that impact on the verification process. Typically, initial verification of claims regarding a particular model involves the testing of one unit (called a Stage I checktest). If the unit fails specific checktesting criteria, 3 more units are tested<sup>2</sup> (called a Stage II checktest). These subsequent tests may be conducted in a different laboratory to the first test.

Differences in the test results represent the outcome of several different types of factors. These can be classified into two general categories:

- (i) random errors, and
- (ii) systematic errors.

These two types of errors are described more fully below.

#### 2.1.1 Random Errors

Random errors are the kinds of errors that are caused by natural variations in materials, human factors, fluctuations in power input etc. Such errors may cause measurements of appliance performance to deviate from the true or “design” performance level. A key feature of random errors is that they are just as likely to be positive as they are to be negative, and over many measurements of performance they average out to be zero.

The main sources of random error in the verification process are:

- **Production Variability**  
All production processes are subject to random fluctuations as a result of manufacturing tolerances, variations in input materials, power fluctuations, human factors etc. These variations in the production process may cause different units of the same model to have slightly different average performance levels. This random error describes the differences in the average performance of different units of the same model due to such production variability.
- **Performance Variability**  
In addition, the same individual unit may perform differently on different occasions under test; eg a pressure switch may terminate fill volume to a different amount each time, even under identical test conditions. Performance variability is often related to the quality of components used in an appliance but it can also be a reflection of the complexity of the process being tested and of the test assessment (e.g. hand soiling of dishes and the subsequent visual assessment of washing performance of a dishwasher). This type of error affects a test’s repeatability.

---

<sup>2</sup> The number of products tested in Stage I and Stage II checktests can vary by product.

- **Random Measurement Error**

If the performance of a single unit is tested twice, in the same laboratory, and using exactly the same equipment and the same staff, then, in addition to the performance variability, there will be some variability in the test results due to random variations in testing equipment, measurement procedure, human factors, etc. This type of error also affects a test's repeatability.

*It is difficult to separate out the error due to performance variability from the random measurement error. Hence the joint impact of these two errors will be referred to as the test repeatability error.*

Only limited information is available on the relative sizes of the errors introduced in the verification measurements due to production variability or test repeatability. More data are available on the combined variability caused by production variability and test repeatability. Table 1 summarises some of the available information in relation to the energy consumption of a number of products in the energy labelling program. These data have been derived from energy labelling applications where 3 different units of the same model are tested in the same laboratory.

**Table 1: Variability in the Measurement of Energy Performance**

<i>Product</i>	<i>Average Standard Deviation<sup>1</sup></i>	<i>Maximum Difference from Average<sup>2</sup></i>	<i>Number of Models</i>
Refrigerator	25 kWh (3.2%)	125 kWh (18%)	1006
Air conditioning (cooling EER)	0.028 (1.0%)	0.25 (9.2%)	2188
Air conditioning (heating COP)	0.033 (1.1%)	0.39 (12.8%)	1469
Clothes Washer (warm wash)	18.9 kWh (4.5%)	102 kWh (23%)	490
Clothes Washer (cold wash)	3.4 kWh (2.9%)	41 kWh (24%)	232
Clothes Dryer	6.1 kWh (1.5%)	40 kWh (6.9%)	114
Dishwasher (cold connect)	8.7 kWh (2.1%)	51 kWh (9.1%)	369

Notes:

1. Sample standard deviation of three units is used to calculate this value. The standard deviations are also measured as a percentage of the sample mean. This is also known as the coefficient of variation. Measuring standard deviations as percentages enables us to make comparisons across different models and different appliance types. The numbers shown are the average of these absolute and relative sample standard deviations across the different models for which three measurements were available for three different units of the same model.
2. The maximum difference is calculated as the absolute difference between the most extreme unit in the sample of three from the mean of the three. In most of these cases (which are generally very unrepresentative) it appears that three very different units were tested (or in some cases one unit was very different from the other two). Specific investigations would be required to ascertain why such large variations were produced in these isolated cases. The minimum difference between all three units was zero for all products.
3. The variance of a series of measurements is obtained by taking the square of the standard deviation.
4. The smaller the standard deviation, the less variability there is in the measurements; ie the more precise the measurements. For example, measurements of the energy performance of clothes washers, with an average standard deviation of 4.5%, are the least precise. However, there are issues which explain the higher standard deviation for this particular product (see below).
5. The calculations were carried out as follows. For each model, test results were available for 3 separate units tested in the same laboratory. The sample mean and sample standard deviation of the energy performance measurements of these 3 units were then calculated and used to obtain the standard deviation as a percentage of the sample mean.

The case of clothes washers above is interesting. The relative variability (maximum difference) of measurements for clothes washers, on average, was the highest of all the products examined (in percentage terms, rather than absolute kWh terms). A typical variability across the three units for any model was around 20 kWh/year. This of course correlated positively to some degree with the amount of energy used by the model, as can be seen in Figure 1, which plots the variability obtained in testing three different units of each clothes washer model against the average amount of energy used by each model in a warm water wash test. But the relative variability of machines with a low annual energy (typically front loading machines) tended to be worse (in percentage terms) than those that used large amounts of energy (typically top loaders that use external hot water).

**Figure 1: Distribution of Absolute Variability for Clothes Washers (Warm Wash)**

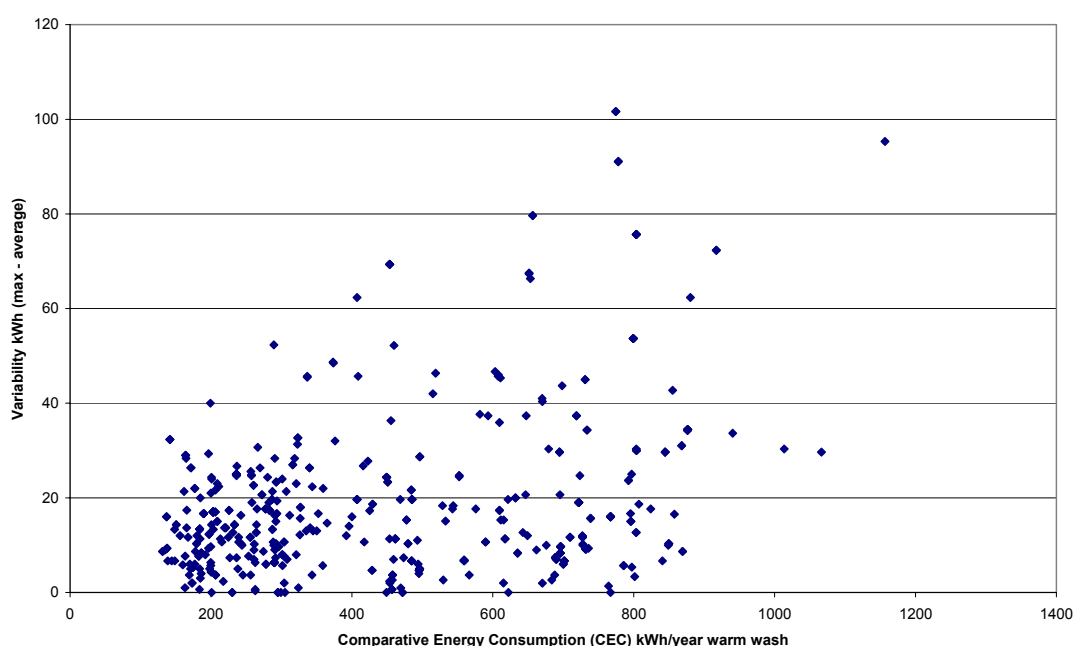


Table 1 refers to the combined error due to production variability and test repeatability. To gain information on the separate impact of these two types of variability, additional tests would be required, namely, several replicate tests conducted on the same unit in the same laboratory. These would provide an estimate of the size of the repeatability error. Typically, estimates of variability are fairly imprecise; hence it's important to eventually gather as much information of this type as possible. However, in reality, little data for replicate tests is readily available.

One good potential source of such data would be the results on clothes washer and dishwasher reference machines, which are required to operate in parallel with test machines under the Australian Standard. However, these data have not been compiled for this report. There are unlikely to be many replicate tests available for other appliance types - if required, such tests may need to be funded. To obtain better data on this aspect, it is recommended that the results of available past replicate tests be collated and subjected to further analysis.

Once an estimate of the combined impact of production variability and repeatability error is available, an estimate of the error due to production variability can be obtained by subtracting the repeatability errors from the combined errors of the type shown in Table 1. More specifically, treating production variability and test repeatability as independent sources of error, the variance of production variability equals the variance of the combined error minus that of the test repeatability error.

### 2.1.2 Systematic Errors

Systematic errors are errors that are not completely random. These errors may have some pattern in them; for example, a series of measurements may have a bias leading to an overstatement (or understatement) of the true performance measure. Such errors can be caused by differences in measuring equipment, calibration errors, differences in procedures between laboratories etc.

For our purposes, the sources of systematic error can be classified into two categories, both of which can be regarded as measurement errors:

- (a) calibration errors, and
- (b) inter-laboratory variability.

- ***Calibration Errors***

Equipment which is not properly calibrated can lead to systematic errors in the measurement of performance levels. Calibration errors cannot be detected in the verification testing program since they are confounded with laboratory-specific factors, such as types of metering equipment, different operating procedures etc. For the present purposes it is assumed that calibration errors are adequately addressed by laboratory accreditation procedures under NATA<sup>3</sup>. For electrical energy consumption, the calibration error is usually less than 2% (typically 1% or better). However, for products such as clothes washers, much of the energy is embodied in hot water drawn into the machine, which means that calibration errors in water temperature measurement and water volume will also contribute to energy errors. For other products such as refrigerators and air conditioners, calibration errors in air temperature measurement can have a large effect on the measurement of performance actually delivered.

- ***Inter-Laboratory Variability***

Performance measurements taken in one laboratory can differ systematically from those taken in another laboratory due to differences in equipment, operating procedures and staff. An estimate of the size of the errors introduced due to inter-laboratory variability can be obtained through a program of round robin tests in which two or more laboratories all carry out tests on the same unit. Estimates of inter-laboratory variability, if established through round robins, will include any calibration errors present.

---

<sup>3</sup> The National Association of Testing Authorities, Australia (NATA) is one of the primary laboratory accreditation agencies in Australia. See [www.nata.asn.au](http://www.nata.asn.au) for details.

## 2.2 Statistical approach to verification

The aim of checktesting is to ensure that manufacturers' energy efficiency and performance claims accurately reflect the information contained within their original application for registration. A failed checktest is generally subject to regulatory action so there needs to be a reasonable degree of certainty regarding the results of the test procedure. Ideally, a unit should only be referred for regulatory action where there is a high degree of certainty that the model has failed one or more of the requirements.

The following sections outline a statistical approach for determining a level of tolerance which results in an acceptable degree of certainty during the verification of manufacture declarations and mandatory performance limits. The statistical model underpinning this discussion is outlined in Appendix A. There are a small number of performance measures and limits where it is not possible to define a tolerance due to the complex nature of the test - these cases are noted in the Administrative Guidelines under Complex Performance Requirements.

### 2.2.1 Verification of manufacturer declarations

Verification of manufacturer declarations is necessary where a performance measure is declared to consumers and where this may be used to compare models. In these cases, consumers need to be reasonably assured that the appliance meets (or is close to) the manufacturer's declaration. The main purpose of a manufacturer declaration is to provide information to the consumer.

Under the energy labelling program, the major declaration for all appliances is **energy consumption** (the Comparative Energy Consumption shown on the energy label), but this discussion also applies to the following types of declarations which may be applicable to MEPS and/or energy labelling:

- Refrigerator and freezer volume
- Air conditioner heating and cooling output capacity
- Clothes washer spin performance
- Lamp lumen output
- Commercial refrigeration total display area

There are a number of other performance measures where the manufacturer declares the level against some standardised measure. These are:

- Clothes dryer capacity
- Clothes washer capacity
- Dishwasher capacity
- Electric hot water delivery capacity (predefined delivery capacities in the standard)
- Electric motor output (shaft) power
- Lamps sizes/ratings that are suitable to be driven by a ballast
- kVA output for distribution transformers

In most cases where the manufacturer declares the performance level in this manner, the manufacturer's declaration is used to determine or set the test parameters in the standard and this is, in practice, verified through the measurement of other performance measures. In the case of hot water delivery capacity, the value declared can be verified through direct measurement.

There are potentially two types of *wrong* conclusion that can be drawn from the results of a sampling and testing process:

1. It could be wrongly concluded that the model doesn't satisfy the manufacturer's declaration, when in fact it does. This type of wrong conclusion is known in statistics as a Type I error.
2. It could be wrongly concluded that the model does satisfy the manufacturer's claim, when in fact it doesn't. This is known as a Type II error.

A Type I error leads to deregistration of a model when, in fact, the manufacturer's declaration is valid. A Type II error leads to a model being allowed on the market when it in fact does not meet the manufacturer's declaration. From a regulatory perspective, Type I errors are more critical in the verification process than Type II errors (although these are still important).

The probability of drawing either type of wrong conclusion depends on the actual average performance level of the model (denoted by " $\mu$ " in Appendix A), and on the variability in product performance (denoted by " $\delta_i + \varepsilon_{ijt}$ ") and in measurement error (denoted by " $\eta_{ijt} + \tau_j + \theta_j$ ").

To gain some insight into the effectiveness of the current verification procedure, consider an example. For simplicity, ignore any systematic measurement errors and any correlation between test results (see Appendix A for further discussion).

The hypothesis to test is that  $\mu = \mu_0$ , where " $\mu_0$ " is the declared performance level against the alternative hypothesis that the performance level is really worse than the declared level. i.e.  $\mu > \mu_0$ . The current test criterion for energy consumption for many products is to reject the null hypothesis if the test performance exceeds the manufacturer's claim by at least 10%.

It is now possible to investigate the types of decision error than can be made. The probabilities of making these errors depend on the real average performance level " $\mu$ ", and on the standard deviations (or, equivalently, the variances) of the various errors.

Table 1 gives some indication of the size of these standard deviations. Typical values for air conditioners and clothes dryers are around 1% to 2%. For clothes washers, refrigerators and dishwashers typical standard deviations are around 3% to 4%. But for individual models the standard deviations can be as high as 20% in the most extreme cases (noting that these cases are very uncommon).

a) **Type I Error: Average performance level is as claimed (or better), but the tests show that the model exceeds the stated claim by at least 10%.**

In the first-stage test of a single unit of the appliance, the probability of reaching such a false conclusion is given by the statistical formula:

$$\begin{aligned} \text{Prob}(\text{Type I error}) &= \text{Prob}[Y_{ijt} > (\mu_0 + 0.1\mu_0); \text{ given the real } \mu = \mu_0] \\ &= \Phi(-0.1/\sigma) \end{aligned}$$

where  $\Phi$  is the standard normal distribution function, and " $\sigma$ " is the relative standard deviation.

For a range of values for the population standard deviation, the probability of a Type I error is shown in Table 2. In this case (and for all subsequent tables in this section), the standard deviation shown is the total expected error from all sources (including any systematic error from laboratory measurements, which could be positive or negative before a laboratory is selected).

**Table 2: Probability of Type I Error When Testing a Single Unit (Stage 1)**

Standard deviation	Prob(Type I error)
1%	0.00%
2%	0.00%
3%	0.04%
4%	0.62%
5%	2.28%
10%	15.87%
15%	25.25%
20%	30.85%

If a model fails the first-stage checktest, 3 new units are normally tested in the second stage checktest. The formula for calculating the probability of reaching a false conclusion in the second stage, given that " $n$ " units are tested in stage 2, is given by the statistical formula:

$$\begin{aligned} \text{Prob}(\text{Type I error}) &= \text{Prob}[Y_{ijt} > (\mu_0 + 0.1\mu_0) \text{ and } Y_{\text{mean}} > (\mu_0 + 0.1\mu_0); \text{ given the real } \mu = \mu_0] \\ &= \Phi(-0.1/\sigma) * \Phi(-0.1 * \sqrt{n}/\sigma). \end{aligned}$$

where  $\Phi$  is the standard normal distribution function, " $\sigma$ " is the relative standard deviation, " $n$ " is the number of units tested at stage 2, and  $Y_{\text{mean}}$  is the average of the stage 2 test results.

Some indicative calculations using this formula with  $n=3$  are shown in Table 3.

**Table 3: Probability of Type I Error After Testing Both Stages 1 and 2**

Standard deviation	Prob(Type I error)
1%	0.00%
2%	0.00%
3%	0.00%
4%	0.00%
5%	0.00%
10%	0.66%
15%	3.13%
20%	5.96%

Note: Since 3 new units are tested in the second stage checktest, the probability is independent of the outcome of the first-stage test. The probability of failing both the first-stage checktest and the second-stage checktest is given by the product of the probabilities of failing each stage separately.

The results in Table 2 and Table 3 show that the probability of a Type I error occurring using the current testing procedure is extremely low -- namely, less than 1 case in 10,000 - provided the standard deviation (from all sources) is about 5% or less. Even with a standard deviation of 10%, the probability of a Type I error after Stage 2 is less than 1 in 100 if the manufacturer's claim is accurate.

For virtually all air conditioners, clothes dryers and dishwashers, the maximum variability is less than about 10% and the current procedure provides a satisfactory probability of Type I error of less than 1%. For refrigerators and clothes washers the maximum variability is at times larger than 10%, although the average standard deviation for these products is less than 4%. Thus, while the current testing procedure provides very low Type I error probabilities (less than 1 in 10,000) for models with typical standard deviations, there will be some isolated models where the variability is larger and hence the probability of a Type I error is larger.

A possible way of overcoming the higher probability of a Type I error in such cases is to introduce a third stage into the testing procedure. The third stage would be invoked if at the second stage the 3 units tested fail the test, i.e. if their average performance exceeds the declared value by more than 10%, and if, in addition, the standard deviation (of these units) is greater than 10% of the mean. In that case it is recommended that three additional units be tested, and the average test performance for these 3 units be used to determine whether the model satisfies the manufacturer's claim. The probability of a type I error for a 3-stage procedure are set out in Table 4.

**Table 4: Probability of Type I Error After 3 Stages of Testing**

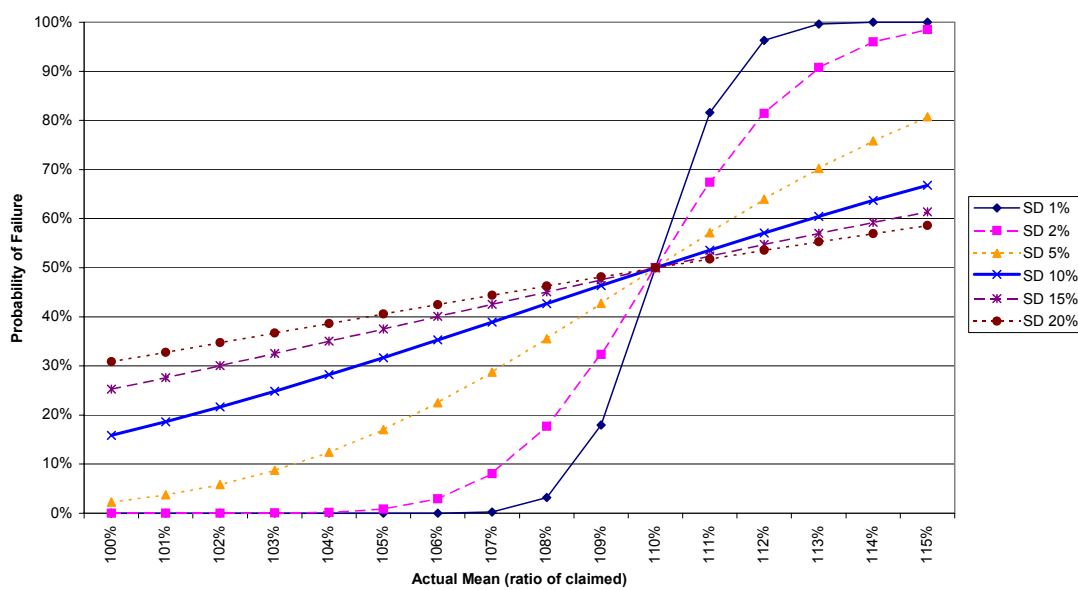
Standard deviation	Prob(Type I error)
5%	0.00%
10%	0.03%
15%	0.39%
20%	1.15%
25%	2.06%
30%	2.93%

Note: The calculation is similar to that used for Table 3, namely, the probability of failing Stage 2 is multiplied by the probability of failing Stage 3, given that the model satisfies the manufacturer's declaration. It's assumed that the sample standard deviation and sample mean are independent, so that the probabilities are not affected by the fact that the Stage 2 sample standard deviation is large.

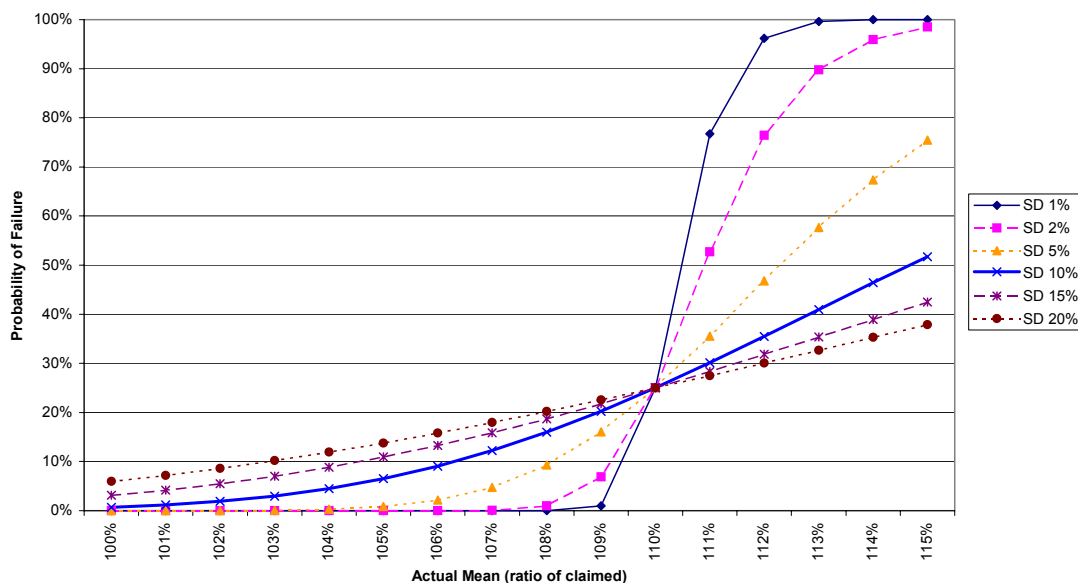
This three-stage procedure produces acceptable Type I error probabilities of just over 1%, provided that the standard deviation is 20% or less. Cases where the standard deviation is larger than 20% are very uncommon, and are likely to be caused by data entry errors, testing errors, defective products or other unusual errors such as product design changes.

However, as the actual production mean of the model starts to exceed the manufacturer's original claim, the probability of a model failing Stage 1 and Stage 2 also increases. The net effect of these two stages is shown in Figure 2 and Figure 3.

**Figure 2: Probability of Failing Stage 1 for Various Standard Deviations**



**Figure 3: Probability of Failing Stage 2 for Various Standard Deviations**



***b) Type II Error: Average performance level is higher (or worse) than that declared by the manufacturer, but the tests show that the model does not exceed the stated claim by more than 10%.***

Type II errors are not as important in verification testing as Type I errors, since they don't result in unfair restrictions on the manufacturer that supplies a product<sup>4</sup>. The probability of a Type II error occurring depends on the real performance level of the model " $\mu$ ", as well as on the standard deviation of the test results.

In the current testing procedure, a Type II error can occur in two ways.

- The model can pass the first-stage test, even though its performance is higher than declared.
- The model can fail the first-stage test, and then pass the second-stage test, even though its performance is higher than declared.

Consider two cases. In the first case, the true performance level of the model is 10% higher (worse) than declared by the manufacturer. In the second case, the performance level is 20% higher (worse) than declared.

***Case 1. Performance level is 10% higher (worse) than the declaration***

The probability of a Type II error occurring at stage 1 in this case is always 50% regardless of the standard deviation. The probability of a Type II error occurring at stage 2 is 25%. Hence the overall probability of a Type II error is 75%.

Since additional stages of checktesting are designed to reduce the probability of a Type I error, they tend to increase the probability of a Type II error. However, this is not seen as critical for checktesting.

***Case 2. Performance level is 20% higher (worse) than the declaration***

The probability of a Type II error occurring at Stage 1 in this case depends on the standard deviation. The statistical formula for this probability is given by:

$$\begin{aligned}
 & \text{Prob}(\text{Type II error; given real } \mu = \mu_0 + 0.2\mu_0) \\
 &= \text{Prob}[Y_{ijt} < (\mu_0 + 0.1\mu_0); \text{ given the real } \mu = \mu_0 + 0.2\mu_0] \\
 &= \Phi(-0.1\mu_0 / 1.2\mu_0\sigma) \\
 &= \Phi(-0.1/1.2\sigma)
 \end{aligned}$$

Indicative calculations are set out in Table 5.

---

<sup>4</sup> However, Type II errors are unfair for competitors who have products which comply with requirements. Reduction of Type II errors ensures a fair competitive market.

**Table 5: Probability of Type II Error When Testing a Single Unit (Stage 1)**

Standard deviation	Prob(Type II error)
1%	0.00%
2%	0.00%
3%	0.27%
4%	1.86%
5%	4.78%
10%	20.23%
15%	28.93%
20%	33.85%

The probability of a Type II error at a Stage 2 checktest presumes that the unit tested at the first stage failed the test, and then the 3 units tested in stage 2 passed the test, even though the real performance level is 20% higher (worse) than declared. The statistical formula for calculating this probability is given by:

$$\begin{aligned} & \text{Prob}(\text{Type II error at stage 2; given real } \mu = \mu_0 + 0.2\mu_0) \\ &= \text{Prob}[Y_{ijt} > (\mu_0 + 0.1\mu_0) \text{ and } Y_{\text{mean}} < (\mu_0 + 0.1\mu_0) ; \text{ given the real } \mu = \mu_0 + 0.2\mu_0] \\ &= \Phi(0.1/1.2\sigma) * \Phi(-0.1 * \sqrt{n}/1.2\sigma). \end{aligned}$$

where  $\Phi$  is the standard normal distribution function, " $\sigma$ " is the relative standard deviation, " $n$ " is the number of units tested at stage 2, and  $Y_{\text{mean}}$  is the average of the stage 2 test results.

The combined probability of a Type II error is the sum of the probability of a Type II error at stage 1, *plus* the probability of a Type II error at stage 2 as given by the above formula. Indicative calculations of the total probability of a Type II error after both stages of testing are set out in Table 6.

**Table 6: Total Probability of Type II Error after Stages 1 and 2 if the Real Performance Level is 20% Higher (Worse) than the Declaration**

Standard deviation	Prob(Type II error)
1%	0.00%
2%	0.00%
3%	0.27%
4%	1.88%
5%	4.96%
10%	26.17%
15%	40.86%
20%	49.41%

Table 6 shows that the probability of a Type II error incurred in the current testing procedure is within acceptable limits provided the standard deviation is less than 5%. For larger standard errors, the probability of a Type II error can be quite high.

Unfortunately, there is no practical way to reduce the probability of a Type II error. The three-stage testing procedure proposed above to reduce Type I error (in which additional units are tested), would only lead to marginal reductions in the probability of a Type II error. The reason is that most of the Type II error occurs at the first stage of the test procedure.

On the other hand, the probability of a Type II error does reduce significantly if the real performance level is much higher than the declaration. As an example, Table 7 shows the probability of a Type II error if it is assumed that the real performance level is 40% higher (worse) than the declaration.

**Table 7: Total Probability of Type II Error after Stages 1 and 2 if the Real Performance Level is 40% Higher (Worse) than the Declaration**

Standard deviation	Prob(Type II error)
1%	0.00%
2%	0.00%
3%	0.00%
4%	0.00%
5%	0.00%
10%	1.61%
15%	7.84%
20%	15.50%

In this case the probability of a Type II error is acceptable, with a value of 5% or less, as long as the standard deviation is less than about 13%. Thus, using the current checktesting approach, the chance that a model which grossly exceeds the declared performance level will pass the checktesting procedure undetected is quite small (except where the standard deviation is very large).

### 2.2.2 Verification that a product meets minimum performance standards

During the verification of a manufacturer declaration, the focus is on verifying that the average performance level of the model is as claimed by the manufacturer. While some units may have a worse performance level than claimed, these can be balanced by units with a better performance level provided the average performance level of the model is as claimed. The main purpose of a manufacturer declaration is to provide information to the consumer.

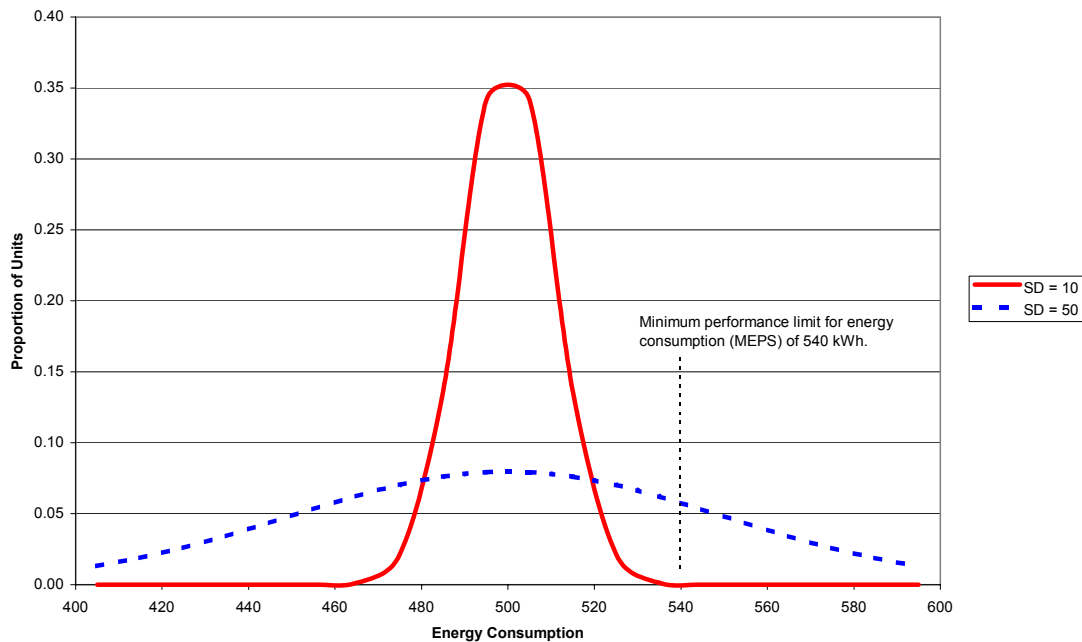
Verification of performance limits (minimum performance standards) has a different objective. In principle, all units of the model should satisfy the performance limit. In practice, product variability might lead to some units of a model which operates close to the performance limit failing to meet the standard. This suggests that the verification of the performance limit should allow for some percentage of failures, say 5% or 10%. The main purpose of a performance limit is to provide a degree of consumer protection (noting that consumers are not normally explicitly informed of performance limits).

Verification of minimum performance standards (MPS) can be represented as a problem in statistical hypothesis testing. The hypothesis to test is that the proportion of units which fail to meet the minimum performance standard (MPS) is "p", against the alternative hypothesis that the proportion is greater than "p". A more formal statement of these hypotheses is given in the statistical Appendix A. Possible values of "p" which seem suitable for the situation under discussion are 5% and 10%. This concept is explained in more detail below.

The proportion of units that exceed a particular value (say a minimum performance standard) obviously depends on both the mean and standard deviation of the performance measure under consideration. Consider the following two hypothetical products shown in Figure 4. Both have a mean energy consumption of 500 kWh/year, but the first has a standard deviation of 10 kWh while the second has a standard deviation of 50 kWh. Consider an example where the allowable MEPS for this product is, say, 540 kWh/year. Literal reading of the standard requires that *no* units should exceed this limit. While visual inspection of Figure 4 indicates that the product with a standard deviation of 10 kWh meets this limit, in fact, more precise calculation shows that 0.003% of units are still expected to fail this MEPS level. For the model with a standard deviation of 50 kWh, some 21% of units are expected to exceed the limit.

Under the practical application of the minimum performance limit proposed above, the proportion of units that fail the MEPS level would be limited to 5% or 10% of the units. In the example shown, the model with a standard deviation of 10kWh would pass, while the model with a standard deviation of 50 kWh would fail.

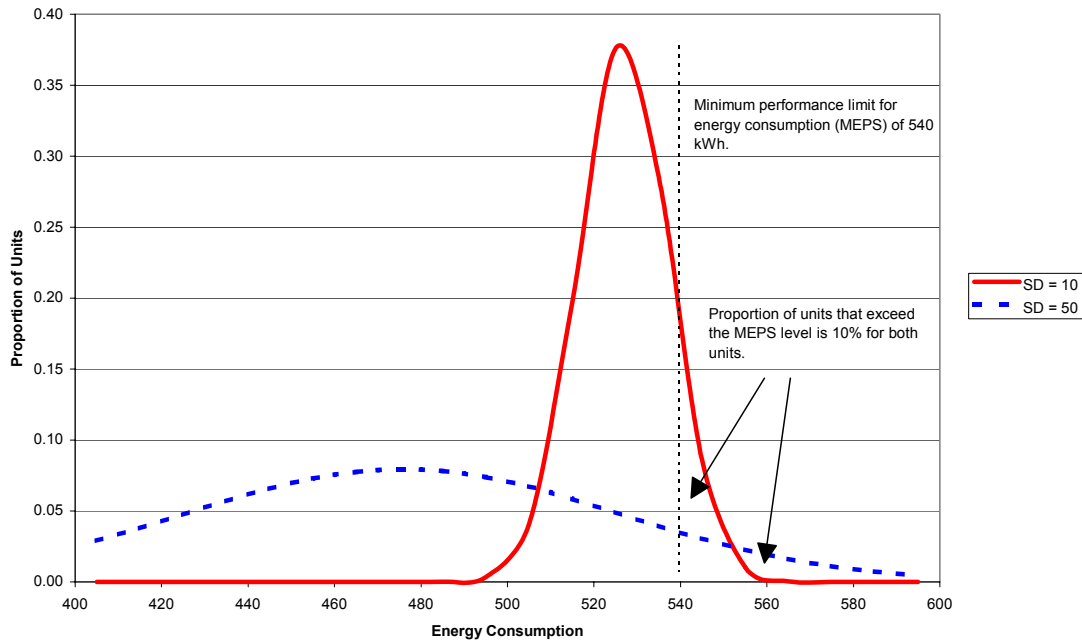
**Figure 4: Energy Distribution of 2 Hypothetical Products**



In the case above, if a proportion of units for each model were allowed to fail the MEPS level, then the model with a 10 kWh standard deviation could in fact have a higher mean energy consumption and still pass the requirement, while the model with the 50 kWh standard deviation would have to have a lower mean energy consumption to meet the MEPS requirement. For 95% of the units to pass the test, the mean + (1.645 × standard deviation) must be less than the MEPS level. For 90% of the units to pass, the mean + (1.2817 × standard deviation) must be less than the MEPS level.

In Figure 5, the maximum allowable mean energy consumption for our two models has been adjusted such that the proportion that is expected to fail the allowable MEPS limit is 10%. In this case the maximum mean allowable energy consumption is 527 kWh/year for the model with a standard deviation of 10 kWh and 476 kWh/year for the model with a standard deviation of 50 kWh.

**Figure 5: Maximum Allowable Mean Energy for 2 Hypothetical Products - 10% fail**



Now consider the probability of drawing wrong conclusions from the test results on a model if such a performance limit were adopted. As before, there are two types of wrong conclusion that can be drawn from the results:

1. It could be wrongly concluded that more than 5% (respectively 10%) of the units of the model fail the performance limit, when in fact this is not true. This type of wrong conclusion is known in statistics as a Type I error.
2. It could be wrongly concluded that less than 5% (respectively 10%) of the units of the model fail the performance limit, when in fact more than 5% (respectively 10%) fail the performance limit. This is known as a Type II error.

As before, a Type I error is of more concern than a Type II error, since a Type I error involves taking action against a manufacturer when this is not warranted. These two types of decision errors are now examined in more detail.

In order to calculate the probability of making Type I and Type II errors, it is necessary to propose a specific sampling/checking procedure. Consider a checktest procedure similar to the current procedure. Typically, in a Stage 1 checktest, one unit of a model is tested. If it fails to meet the MPS, a Stage 2 checktest is invoked in

which another 3 units are tested. The question is what criterion to use at Stage 2 to reject the model. The possibilities are to reject the model if:

- (a) at least one of the three extra units fail to meet the MPS
- (b) at least two of the three extra units fail to meet the MPS
- (c) all three extra units fail to meet the MPS

**a) Type I Error.**

Suppose that the model satisfies the criterion that no more than a proportion of "p" of units fail the minimum performance standard. If one unit is tested then the probability of drawing the wrong conclusion that the model doesn't satisfy the criterion is "p". This assumes that there is no measurement error.

The situation is more complicated if there is measurement error. Measurement error increases the observed proportion of units failing to meet the MPS above the nominal proportion "p", which leads to an increase in the probability of a Type I error. The amount by which "p" is inflated depends on how large the standard deviation of measurement error is compared to the standard deviation of performance variability. Table 8 provides indicative calculations of the relationship between the nominal "p" and the observed, inflated proportion that needs to be used to calculate Type I errors. The observed proportion is denoted by "p\*". The statistical formula used for calculating "p\*" is given by:

$$p^* = \text{observed proportion} = \Phi[\Phi^{-1}(p)/\sqrt{(1+r^2)}]$$

where "Φ" is the standard normal distribution function, "Φ<sup>-1</sup>" is the inverse standard of the normal distribution function, "p" is the nominal proportion of units failing the MPS, and "r" is the standard deviation of the measurement error as a percentage of standard deviation of true performance variability.

**Table 8: Impact of Measurement Error on Proportion of Units Failing MPS**

Std. Dev. of Measurement Error as a Percentage of Std. Dev. of True Performance Variability "r"	Nominal "p" =5%	Nominal "p" =10%
	<i>p*</i>	<i>p*</i>
10%	5.08%	10.11%
25%	5.53%	10.69%
50%	7.06%	12.58%
75%	9.41%	15.26%
100%	12.24%	18.24%

Note: The nominal proportion "p" refers to the case where there is no measurement error. Hence it is the true proportion of units that fail to meet the MPS. The percentages in the body of the table are the proportion of failures expected to be observed in test results if there is measurement error. These are the proportions that need to be used in the binomial probability formula to calculate the probability of Type I errors.

Table 8 shows that if measurement error is fairly small, with a standard deviation which is only 10% of the standard deviation in the true performance of different units, then the observed proportion of units which fail the MPS is only slightly higher than the nominal proportion. However, as the measurement error increases, the discrepancy between the nominal "p" and the proportion of units which actually fail the MPS in laboratory tests increases accordingly.

Suppose that the one unit tested at Stage 1 fails to meet the MPS. The current procedure would be to test another 3 units. The probability that at least one, two or all three of these units fail to meet the MPS, when the observed proportion is "p\*", is given by the cumulative binomial probability formula, multiplied by the probability of failing the Stage 1 test, ie. by "p\*". These probabilities are set out in Table 9.

**Table 9: Probability of Type I Error**

Number of Units Failing the Stage 2 Test	p*=5%	p*=10%	p*=13%	p*=18.24%
1 or more	0.71%	2.71%	4.44%	8.27%
2 or more	0.04%	0.28%	0.60%	1.60%
3 or more	0.00%	0.01%	0.03%	0.11%
New 3-stage procedure	0.04%	0.28%	0.60%	1.60%

Suppose that the one unit tested at Stage 1 fails to meet the MPS. The current procedure would be to test another 3 units in Stage 2. Let p\* be the effective proportion of failures for a particular model, allowing for measurement error. Then Table 9 sets out the probabilities that a model fails Stage 1, and that in Stage 2 at least one, two or all three of the 3 units tested fail to meet the MPS. These probabilities are calculated using the cumulative binomial probability formula, multiplied by p\* to capture the probability of failing Stage 1.

Suppose the value of "p" is chosen to be 10%. That is, based on true performance without any measurement error, up to 10% of units of the model are permitted to fail the MPS (see Appendix A for a description of measurement error). With measurement error, our measurement of the performance level is not exact, and this will increase the proportion of units of this model that can be expected to fail the MPS in practice. Assume an extreme case where the measurement error has a standard deviation that is of the same magnitude as the standard deviation of the variability of the true performance across different units of a model. Then Table 8 shows that a "no measurement error" failure rate of p=10% corresponds to an expected observed failure rate of p\*=18.24%. However, from Table 9 we can see that, even in this extreme case, if we were to adopt the criterion that a model be rejected if at Stage 2, two or more of the three units tested fail to meet the MPS, then the probability of a Type I error is still only 1.6%.

If the measurement error is smaller, then the probability of a Type I error is also smaller. However, even if there were no measurement error, ie. if p\* = p =10%, then the same test criterion seems appropriate. Tightening the criterion so that the model is rejected as soon as one or more Stage 2 units fail to meet the MPS, would lead to a Type I error probability of 2.71% which is probably unacceptably high.

At this stage there is little information available from which to calculate the relative contributions of measurement error versus performance variability. However, the examples illustrated above show that the suggested test procedures for the  $p=5\%$  case and the  $p=10\%$  case are quite robust with respect to measurement error, in the sense that they still deliver very reasonable probabilities of a Type I error with measurement error ratios as high as 50%.

Although the error associated with reproducibility is random in nature (a laboratory can be either high or low), once a laboratory is selected for checktesting, the actual offset for a performance reading (high or low) becomes quite critical when compared to the minimum performance limit. It is proposed to undertake further analysis of this aspect.

## **b) Type II Error**

Suppose the model does not satisfy the criterion that at most 5% (respectively 10%) of units should be allowed to fail the minimum performance standard. It might still be wrongly concluded that the model satisfies the criterion, ie that less than 5% (respectively 10%) of units fail the performance standard.

The probability of drawing the wrong conclusion depends on what proportion of units actually fail to meet the minimum performance standard, ie. on the true value of "p".

Assume that  $p=20\%$ . Then the probability that the one unit tested at Stage 1 satisfies the MPS is 80%. This would be a Type II error. If the Stage 1 unit fails to meet the MPS, then it still possible that the model is accepted at Stage 2. This is again a Type II error. The probability of this happening depends on which rejection criterion is used. If the model is rejected when at least one unit fails to meet the MPS, then the combined Stage 1 and Stage 2 probability of a Type II error is 90.2%. If the alternative criterion of rejecting the model when at least two Stage 2 units fail to meet the MPS is used then the combined Stage 1 and Stage 2 probability of a Type II error is 97.9%.

These results suggest that the proposed procedure has almost no power to detect whether a model fails to satisfy the MPS. However, this situation only arises because in practical terms a value of  $p=20\%$  is not very far away from the allowed proportion of failures. It is likely that if production tolerances are quite tight, then when a model fails to meet the MPS, it will fail the MPS by a considerable margin, ie. one would expect almost all units of the model to fail the MPS. If the above calculations are repeated with  $p=0.95$ , the probability of a Type II error is 5.01% for the first rejection criterion (ie at least one unit fails at Stage 2), and a probability of 5.61% for the second rejection criterion (ie at least two units fail at Stage 2). These are quite acceptable probabilities for Type II errors.

Unfortunately, the probability of a Type II error only reduces slowly as more units of a model are tested. Hence testing more units to reduce the probability of a Type II error is impractical. An common alternative approach to reducing the Type II error is

to accept a higher probability of a Type I error. However, in the present case there is little scope for doing that, since effectively this would lead to rejecting the model at Stage 2 regardless of the test outcomes.

The above analysis of the probability of a Type II error has assumed that there is no measurement error. The effect of measurement error on the probability of Type II errors depends on the true value of "p". If this value is close to the null hypothesis value of 5% or 10% then the probability of a Type II error is lower than in the above example. Thus, the Type II error probabilities in the example above, where it is assumed that  $p=0.2$  will be reduced somewhat. On the other hand, if the true "p" is much larger than 5% or 10%, then the probability of a Type II error is increased.

### **2.2.3 Performance measures where a tolerance cannot be defined**

For complex tests with multiple pass/fail criteria, it is not possible to define simple validity criteria. There are several examples that fall into this category including:

- Refrigerator temperature operation test - the unit is tested with one to several control setting combinations to establish that the unit has at least one combination of control settings where all internal compartment temperature requirements can be met simultaneously under the specific ambient temperature (tests at 10°C, 32°C and 43°C are required).
- Dry clothes in a single operation - clothes dryers are required to reach the specified final moisture content (6% of bone dry mass) within a single control setting (either via a timer or with an automatic sensor) before cooldown commences. There is no continuous measure for this test for which to define a tolerance (the dryer either passes or fails).
- Clothes dryer maximum allowable drum temperature - theoretically the temperature measurement for this test is on a continuous scale so a tolerance could be defined. However, in practice, the temperature recording strips used record temperatures at about 5°C intervals and strips are selected to record temperatures just below the allowable maximum temperature (typically 128°C to 130°C) with the next step well above the allowable temperature. In effect, the tolerance for this test is built into the type of equipment used.
- Air conditioner maximum operation test - this involves running the air conditioner under extreme conditions then shutting the power off. The air conditioner must meet various requirements after the restoration of power including no motor or wiring damage, to commence operation again within a specified period after the restoration of power without overload trips and to operate continuously for 1 hour.
- Hot water delivery capacity - this test measures the volume of hot water that can be delivered at a specified flow rate until a temperature drop of 12°C is reached (it is effectively a measure of how well the cold water is stratified when it displaces hot water in the storage tank). It is difficult to define a tolerance for this test as flow rate, volume and temperature in combination all define the performance measure. Hot water delivery capacity is not a performance limit, but the value declared by the manufacture can be verified through testing.

It is recommended that the above performance measures and limits be verified or applied as defined in the relevant standards without any tolerance values.

### **3. Principles for Verification**

There are a number of cases where small changes to standards have been implemented over the years, so it is important that the version of the standard used for the original application for registration also be used for verification. For example, the original regulations for clothes dryers specified an initial moisture content of 100% of the bone dry mass, whereas the 1996 standard specifies 90% moisture. It is critical to know which version of the test method was used to determine the energy label value before a clothes dryer energy or other performance measure can be verified through a check test.

#### **3.1 Verification of Manufacturer Declarations**

Under the current system, a single unit is tested and if this fails the screen test criterion for energy (currently 110%), a further 3 units are tested. If the average of these 3 units exceeds 110% of the manufacturers claim, the model is deemed to have failed.

Given the general analysis in this discussion paper, the use of this process for the validation of energy consumption appears to be sound and reasonable. For typical levels of variability and error (from the limited available) there is only an extremely small chance of a deregistration proceeding if this verification procedure is used on a model for which the manufacturer's original declaration is in fact correct. Of course, as the actual mean value for a model increases above the manufacturer's claim, the chance of failing the current verification procedure also increases.

In cases where the sample standard deviation for the 3 units of a model tested in Stage 2 exceeds 10%, consideration could be given to testing more units to minimise the possibility of an incorrect decision, ie. to introducing a third stage into the testing procedure. However, the cases where this would apply would be extremely limited.

It is recommended that the "+10%" rule as current specified for energy also be applied to the following performance measures (which are measured on a continuous scale):

- Air conditioner heating and cooling capacity
- Air conditioner EER and COP
- Clothes washer spin performance

It is recommended that the case of refrigerator volume measurements be treated slightly differently to other performance declarations. The volume of a refrigerator compartment is determined at the time of manufacture. Given that reasonably sophisticated equipment is used in the production of appliances, it could be expected that the variability in internal volume between units of the same model from the same production line are likely to be quite small (probably less than 1%). The measurement of volume, while subject to some potential errors (eg for curved or unusual shaped liners), is not an operational measurement (like energy or water

consumption) and the actual volume will not change in time or space (accepting that there may be very small changes as a result of temperature variation). Generally speaking measurements of length can be made with some certainty. The refrigerator standard AS/NZS4474.1 allows a 3% (or 1 litre) tolerance on claimed volumes versus measured volumes for each compartment and a 3% (or 5 litre) tolerance overall volume. Given that there may be some measurement error and that under normal circumstances only a single unit will be checked for volume, some account of the possible measurement error needs to be explicitly incorporated into the allowable tolerance. Given that the measurement error may be as high as 2% in difficult situations, it is recommended that a 5% tolerance for refrigerator compartment volume be allowed during check testing.

It is recommended that the allowable limits for declared performance measures for new products introduced into the program should be considered by NAEEEC on a case by case basis and these should be included in the Administrative Guidelines.

### **3.2 Verification of Performance Limits**

For the verification of minimum performance limits, it is assumed that the actual performance across individual units of the same model is normally distributed. The standard generally specifies that each unit shall meet the required performance limit, where these are set.

Under a normal distribution, it is not possible to be assured that all units will be able to pass the standard minimum requirements. For the verification of minimum performance limits, it is suggested that a practical requirement would be to allow the worst 10% of units of a particular model to fail the limit stated in the standard (meaning that 90% are required to pass the limit). If it is assumed that the measurement error is equal to the variability of the test measurement (based on the limited data available), during a check test it would be reasonable to allow about 18% of units to fail the requirement.

In a practical terms it is proposed that if a unit is check tested and fails a performance requirement, a further two units be tested. If the additional two units pass the requirements, then the model is deemed to have complied. If the two additional units fail the requirements, the model is deemed to have failed. If one passes and one fails, it may be necessary to conduct a test on one more unit to determine the outcome.

The issue of the nature and magnitude of reproducibility (measurement) errors between laboratories is still unresolved at this stage and will obviously have some impact on the verification of a minimum performance limit. It is proposed to undertake further analysis of this aspect.

### **3.3 Other Recommendations**

It may be necessary to refine the approaches above once more data regarding the level of variability and error has been collected and as new products are introduced into the program. Inter-laboratory errors in particular could have a significant influence on the recommended procedures. In particular, the following types of information should be collected where possible:

- Replicate test data to establish test repeatability;
- Routine collection of data on 3 units from energy labelling registrations (and other performance measures where possible);
- Analysis of round robin test results to establish inter-laboratory error (test reproducibility).

Contact Lloyd Harrington of EES for further information or comments.  
Tel 03 56266333 Fax 03 56266442 Email: [lloydh@ozemail.com.au](mailto:lloydh@ozemail.com.au)

## **Appendix A: A Statistical Model for Verification Testing**

The test performance of an appliance can be represented by the following equation:

$$Y_{ijt} = \mu + \delta_i + \varepsilon_{ijt} + \eta_{ijt} + \tau_j + \theta_j \quad (1)$$

where

$Y_{ijt}$  = the test measurement for unit "i" in laboratory "j" on test occasion "t".

$\mu$  = the actual average performance level for the model. This is a fixed, unknown number, which represents the average performance over many units of the model, assuming there is no measurement error of any kind.

$\delta_i$  = the average deviation of the particular unit from the average performance level for the model. This deviation is a random error due to production variability.

$\varepsilon_{ijt}$  = the random deviation from the unit's average performance level on test or operating occasion "t". This deviation is a random error due to operating variability.

$\eta_{ijt}$  = the random error occurring on test occasion "t" as a result of random measurement errors due to human factors, random variations in the metering equipment etc.

$\tau_j$  = the systematic calibration error for the measurement equipment used in laboratory "j".

$\theta_j$  = the systematic error in laboratory "j" due to operating procedures etc.

All the terms above, except  $Y_{ijt}$  and " $\mu$ ", are random errors with zero mean and zero correlation with all other errors.  $Y_{ijt}$  is a random variable with mean " $\mu$ ", while " $\mu$ " itself is an unknown constant. In the absence of information on the distribution of the various errors, it is assumed that all the errors have normal distributions.

Test measurements for the same model will vary from test to test. Part of the difference is due to the fact that performance is actually different. Factors that may contribute to differences in performance are:

- if a different unit is tested (captured by " $\delta_i$ ")
- random differences in performance due to operating variability; eg. a difference in the temperature at which a valve opens (captured by " $\varepsilon_{ijt}$ ").

The remainder of the difference in test measurements is due to differences in various measurement errors (captured by " $\eta_{ijt}$ ", " $\tau_j$ " and " $\theta_j$ ").

The total variability in test measurements around the mean performance level " $\mu$ ", can be characterised by the variance of  $Y_{ijt}$ , say " $\sigma^2$ ". This variance can be decomposed into the contributions made by all the different random components shown in equation (1), i.e.:

$$\sigma^2 = \sigma^2(\delta) + \sigma^2(\varepsilon) + \sigma^2(\eta) + \sigma^2(\tau) + \sigma^2(\theta) \quad (2)$$

Notes:

- If the same unit is tested in two different tests in different laboratories, then  $\delta_i$  will be the same on both occasions. This introduces a correlation between the test measurements equal to  $\sigma^2(\delta)/\sigma^2$ .
- Similarly, if two tests on different units are carried out in the same laboratory then  $\tau_j$  and  $\theta_j$  will be the same on both occasions. This introduces a correlation equal to  $[\sigma^2(\tau) + \sigma^2(\theta)]/\sigma^2$ .
- If the same unit is tested twice in the same laboratory, then the correlation between the test results is  $[\sigma^2(\delta) + \sigma^2(\tau) + \sigma^2(\theta)]/\sigma^2$ .
- In practice, the variability introduced by  $\varepsilon_{ijt}$  and  $\eta_{ijt}$  cannot be separately identified.

The 5 variance terms on the right hand side of (2) can be split into two groups:

- terms relating to variability in actual performance;
- terms relating to various measurement errors.

If there were no measurement errors of any type, then the test measurements would only be subject to the variability introduced by " $\delta_i$ " (the average deviation of a particular unit from the average performance level for the model), and " $\varepsilon_{ijt}$ " (the random deviation due to operating conditions on test occasion "t" leading to a performance above or below the particular unit's average performance level). Thus, if there are no measurement errors, the actual performance of the model is appropriately characterised by a normal distribution with variance:

$$\sigma^2(\text{performance}) = \sigma^2(\delta) + \sigma^2(\varepsilon) \quad (3)$$

The combined effect of the various measurement errors can be characterised by a normal distribution with variance:

$$\sigma^2(\text{measurement error}) = \sigma^2(\eta) + \sigma^2(\tau) + \sigma^2(\theta) \quad (4)$$

The variance of the test measurements is the sum of these two main types of variance, i.e.:

$$\sigma^2 = \sigma^2(\text{performance}) + \sigma^2(\text{measurement error}) \quad (5)$$

Thus, the variability of the test measurements will be larger than the variability of actual performance scores because of the effect of the measurement variance term in (5).

### ***Verification of Manufacturer Declarations***

Verification of manufacturer declarations is essentially a test of whether the average performance of different units of the model is equal to, or better than that claimed by the manufacturer. This is equivalent to testing a hypothesis about  $\mu$ . Suppose that the manufacturer claims that energy consumption for the model is no more than  $\mu_0$ . Then verification amounts to testing the null hypothesis:

$$H_0: \mu = \mu_0$$

against the alternative hypothesis that energy consumption really exceeds the declared level. i.e.:

$$H_1: \mu > \mu_0$$

We can test these hypotheses using standard statistical testing procedures. In the absence of adequate data to determine the distribution of the test measurements, we can assume that the test measurements are normally distributed.

The presence of measurement errors increases the probability of falsely rejecting the manufacturer declaration when in fact it is true. To reduce the possibility of this occurring, it is prudent to adopt a stringent significance level of say 1%.

The true significance level in terms of actual performance, i.e. using  $\sigma^2(\text{performance})$  rather than  $\sigma^2$ , will be higher than the nominal significance level.

### ***Verification of Minimum Performance Requirements***

Minimum performance standards ideally require all units of the model to meet the standard. However, for models performing close to the standard, it is likely that some proportion of units will fail to meet the standard. Suppose that we allow a proportion of "p" of the units of a model to fail the minimum standard. Then the statistical criterion we wish to check can be written as the null hypothesis:

$$H_0: \text{probability}[(\mu + \delta_i + \varepsilon_{ijt}) > \text{maximum consumption}] = p$$

ie. the proportion of units of the model which fail the standard equals "p", against the alternative hypothesis:

$$H_1: \text{probability}[(\mu + \delta_i + \varepsilon_{ijt}) > \text{maximum consumption}] > p$$

We cannot observe  $(\mu + \delta_i + \varepsilon_{ijt})$  directly. Instead the test measurements we observe also include various measurement errors given by  $(\eta_{ijt} + \tau_j + \theta_j)$ .

The variability increase resulting from these measurement errors will increase the observed proportion of test results that fail to meet the performance standard above "p" when the true proportion is actually "p". The amount by which the observed proportion will exceed "p" depends on the relative size of the variance terms. Until such time as more is known about the relative importance of the various sources of error, it is prudent to set a level for "p" which is not too stringent, say 5% or 10%.