Ozzy.med@gmail.com

A Rational Basis for Medical Devices Design Verification

Purpose

The purpose of this document is to consolidate a thought process, online research results and some insights, and to provide **a rationale-set** and **a method** for use in medical devices design verification.

Important notes:

1. I researched the topic and prepared this summary with specific focus on medical devices. This field has unique characteristics, and thus my definitions, assumptions, insights and methods are not necessarily transferable to any other field. This summary is intended to be useful for professionals dealing with medical devices design and development.

2. Medical devices Design Validation is a separate topic, not to be mixed with Design Verification. A separate article will address Design Validation.

Definitions and abbreviations

Confidence Level	$(1 - \alpha)$ The probability that a specific statistical prediction about a specific population will be supported (show true) in a random sample drawn from that population; or Given a large-enough number of random sample-sets, the proportion of sets that will maintain the prediction
Coverage	The fraction (or proportion – P) of items that actually fall within some specific numerical limits
D&D	Design and Development
GR&R	Gauge R&R / Gauge Repeatability and Reproducibility
MD	Medical Device(s)
QSR	Quality System Regulation (USA 21 CFR part 820)

References

[1] ISO 13485:2016 – *Medical devices -- Quality management systems -- Requirements for regulatory purposes*

[2] 21 CFR 820 – Code of Federal (USA) Regulations, Title 21 (Food and Drugs), Chapter I, Subchapter H (Medical Devices), part 820: *Quality System Regulation*

[3] ISO 9000:2015 – *Quality management systems - Fundamentals and vocabulary*

[4] 2006 ASQ/ASA Fall Technical Conference / Manuscript No. 189, *An Honest Gauge R&R Study*, Donald J. Wheeler (January 2009 Revision)

[5] *The Secret Foundation of Statistical Inference*, Donald J. Wheeler, www.qualitydigest.com, December 2015

[6] *Rational Sampling*, Donald J. Wheeler, www.qualitydigest.com, July 2015

[7] *Economic Control of Quality of Manufactured Product*, Walter Shewhart (ASQ, rerpint 2002)

[8] *Understanding Variation: The Key to Managing Chaos*, Donald J. Wheeler (SPC Press Inc., 2000)

[9] *Statistical Tolerance Intervals*, Donald J. Wheeler, www.qualitydigest.com, January 2016

[10] *Tables for Normal Tolerance Limits, Sampling Plans, and Screening*, Robert E. Odeh and Donald B. Owen (Marcel Dekker, Inc., 1980)

[11] ISO 16269-6:2014 – *Statistical interpretation of data — Part 6: Determination of statistical tolerance intervals*

What does Medical Devices Design Verification entail?

According to the Cambridge Dictionary, *Verification* (in general) is an act of proving or checking that something exists or is true or correct. So literally, *Design Verification* would be an act of checking that a design is "true" or "correct". This definition is a too vague to be of practical use, and a more targeted definition is necessary.

Medical devices development and manufacturing are strongly governed by regulatory / standardized definitions and requirements. These are typically country- or region-specific; however, as of writing this article (August 2019), ISO 13485[1] already represents the current best methodology for MD realization in most of the leading markets. The most notable exception is the USA, where the governing methodology is currently prescribed in 21 CFR part 820[2]. However, the FDA has announced in 2018 that it intends to transition from the existing QSR to an adoption of ISO 13485, and to begin this process in late 2019. Hence, ISO 13485's definitions, concepts and requirements seem very pertinent – almost regardless of market – for most intents and purposes related to MD D&D.

In terms of standardized MD development methodology, *Design Verification* falls in the realm of *Design Control* ("Design and Development" in ISO 13485's terms).

ISO 13485:2016, s. 7.3.6 *Design and Development Verification* reads:

> **7.3.6 Design and development verification**
>
> Design and development verification shall be performed in accordance with planned and documented arrangements to ensure that the design and development outputs have met the design and development input requirements.
>
> The organization shall document verification plans that include methods, acceptance criteria and, as appropriate, statistical techniques with rationale for sample size.
>
> If the intended use requires that the medical device be connected to, or have an interface with, other medical device(s), verification shall include confirmation that the design outputs meet design inputs when so connected or interfaced.
>
> Records of the results and conclusions of the verification and necessary actions shall be maintained (see 4.2.4 and 4.2.5).

According to this clause, the purpose of Design Verification is to **ensure** (or confirm) that the Design Outputs have met the Design Input requirements. But **how** should that be **ensured**? The term "ensure" indicates that a resulting high level of confidence is required. The above clause provides some clues, but it's not specific enough to be of direct practical use.

- The methods need to be prescribed upfront, but what methods are appropriate?...
- The acceptance criteria need to be prescribed upfront, but which are appropriate?...

- What statistical techniques (if any) are appropriate for inclusion in Design Verification plans, and how can sample sizes be rationalized?...
- Results and conclusions need to be recorded, but how should conclusions be drawn?...

The following is an attempt to address these questions.

A closer look into definitions may help clarify what's required in order to ensure that the Design Outputs have met the Design Input requirements. S. 3 *Terms and definitions* refers to ISO 9000:2015 for terms and definitions not given in ISO 13485:2016 itself. The terms and definitions given in ISO 13485:2016 s. 3 are not helpful for interpreting s. 7.3.6, so a look into ISO 9000:2015 is necessary.

The ISO 9000:2015 definition for *Verification* is given in section 3.8.12. Parsing that definition, substituting the relevant terms with their own definitions given in ISO 9000:2015, and rearranging, in the context of ISO 13485:2016 s. 7.3.6 (MD Design Verification), may yield the following, more useful definition:

***Verification**:*
Provision of information, the trueness/correctness of which can be checked, obtained through either
- *observation,*
- *alternative calculation,*
- *reviewing documents,*
- *inspection,*
- *measurement,*
- *testing, or*
- *other means;*

confirming that specified requirements have been fulfilled.

ISO 9000:2015 describes the D&D process as a process of furthering / elaborating / refining requirements, for example from what the product needs to do and how it needs to do it (functional and performance requirements) to what dimensions its various components need to have and what types of raw materials are required for making them (technical specifications and drawings). Hence, the output of this process (the Design Output) is essentially a set of requirements, conveyed in a set of documents. Prototypes / physical product builds are sometimes referred to as part of the Design Output; however, this is mostly inconsistent with the above approach.

The important point is that the Design Output is not a physical object; and therefore, by default, answering the question of whether it fulfils the input requirements should be **a theoretical / analytical exercise**. Inspection of physical objects, measurement and testing should only be a last resort, where review and theoretical/computational analysis are too complex or fail to provide a high level of confidence ("ensure", as in ISO 13485:2016 s. 7.3.6).

This is a very important notion, because it allows eliminating a lot of noise from Design Verification activities. In a sense, the design being verified is a "pure", invariant entity. This is of course never the case for a physical product, where production processes, use environment and countless factors introduce a lot of variation that should be accounted for if the product is to be properly evaluated against requirements. This is not to say that the design is always nominal; it may include all kinds of tolerances that once called out must be accounted for. However, the design shouldn't have any "noise"

and where no tolerance is specified (either custom or default) the design should be assumed nominal; but this should be used sparingly, and only for good reasons. Where uncertainty remains about the design, it should generally be seen as "under-defined", and the issue resolved before attempting verification.

In most cases verifying a requirement through review or theoretical analysis will be more efficient (i.e. require less resources) and provide a higher level of certainty. So instead of defaulting to evaluation of physical objects, get used to asking yourself: Can I somehow provide other objective evidence that the "pure" design itself meets the Design Input requirements?

When there's no option but to measure

The most fundamental layer of verification uncertainty is in the measurement itself. To be able to do anything meaningful with results obtained from measurements, the measurement process itself must be characterized and shown to be fit for purpose.

Testing also falls under measurement. While it is possible to conduct binomial testing (go/no-go), it's much more efficient (i.e. a better use of resources), to aim for continuous data – numerical results that can be placed along a scale or an axis – where practicable. In statistical terms, the latter provide much more insight than the former, per the same sample size.

Unfortunately, any meaningful measurement process characterization is very specific. It applies to a certain device design, a certain characteristic of that device, certain measurement equipment, certain operators (unless shown otherwise), a certain measurement environment (e.g. lighting, temperature etc.), certain procedures and instructions, and so on. So, the first step in any meaningful measurement or testing is the qualification of that specific measurement or testing. If the results obtained in this step indicate that the measurement process doesn't introduce too much uncertainty, the same data can be used for the verification itself.

In his article ***An Honest Gauge R&R Study***, Dr. Wheeler rationalizes and prescribes a method for Gauge R&R Study that makes a lot of sense to me, with a bonus of detailing a mathematical / statistical rationale[4]. For practicality, that method is reproduced here, somewhat adapted to suit the current context and needs; however, it's strongly recommended to read the entire original article in order to get the background and a better understanding of the topic.

DEFINITIONS
o:: number of operators participating in the GR&R study (min. 2, max. 5)
p:: number of parts (physical objects) being measured in the GR&R study (min. 2; recommended 5)
n:: number of times each part will be measured in the GR&R study (n = 2, 3, or 4)
For simplicity, make sure p*n and o*n don't exceed 15.

I. DATA COLLECTION
Let o operators measure each of p parts n times each.
All parts should come from a production process representative of the current D&D state, which corresponds with the design being verified. For allowing meaningful statistical inference, the goal is sampling all p parts from an apparently ***Homogenous Process***, i.e. sampling parts manufactured in

sequence (or close), under settings/conditions not expected to vary greatly throughout the production run, as well as from run to run in the near future. The sampled parts should be marked such that the time-order of their production remains available for subsequent analysis.
Recommendation: Have each operator measure all p parts on n different days, once per day; ensure results obtained on previous days or by other operators are not available to the operators during measurement (i.e. record results on a new sheet every time), and alter the parts measurement order (shuffle) each day.
Arrange the obtained data into k subgroups, each containing the n results for a specific part measured by a specific operator (k = o*p).
Compute the range (= max. - min.) for each subgroup.
Compute the Average Range for the k subgroups, and multiply by D_4 (a function of n, see below) to find the Upper Range Limit.

n	2	3	4
D_4	3.267	2.574	2.282

If any of the subgroup ranges exceed this upper limit you need to find out why.

II. ESTIMATES OF VARIANCE COMPONENTS
Use the Average Range from above to estimate the Repeatability Variance Component (the equipment's contribution to variation in the measurement results):

$$\hat{\sigma}_{pe}^2 = \left[\frac{\bar{R}}{d_2} \right]^2$$

n	2	3	4
d_2	1.128	1.693	2.059

Next, estimate the Probable Error: $0.675 \sqrt{\hat{\sigma}_{pe}^2}$
If the measuring system in use has an Effective Measurement Increment > 2 Probable Errors, its resolution is too low and this needs to be addressed before proceeding.

Next, compute the Operator Averages (a set of o values) and use the range of these o values (R_o) to estimate the Reproducibility Variance Component (the operators' contribution to variation in the measurement results):

$$\hat{\sigma}_o^2 = \left\{ \left[\frac{R_o}{d_2^*} \right]^2 - \frac{o}{n\,o\,p}\, \hat{\sigma}_{pe}^2 \right\}$$

p*n =	4	6	8	9	10	12	15
d_2* for o = 2	2.151	2.604	2.906	3.025	3.129	3.305	3.513
d_2* for o = 3	2.120	2.581	2.886	3.006	3.112	3.289	3.499
d_2* for o = 4	2.105	2.570	2.877	2.997	3.103	3.282	3.492
d_2* for o = 5	2.096	2.563	2.871	2.992	3.098	3.277	3.488

Next, add the estimates above to get the estimated Combined R&R Variance Component:

$$\hat{\sigma}_e^2 = \hat{\sigma}_{pe}^2 + \hat{\sigma}_o^2$$

Next, compute the Part Averages (a set of p values) and use the range of these p values (R_p) to estimate the Product Variance Component:

$$\hat{\sigma}_p^2 = \left[\frac{R_p}{d_2^*}\right]^2$$

o*n =	4	6	8	9	10	12	15
d_2* for p = 2	2.151	2.604	2.906	3.025	3.129	3.305	3.513
d_2* for p = 3	2.120	2.581	2.886	3.006	3.112	3.289	3.499
d_2* for p = 4	2.105	2.570	2.877	2.997	3.103	3.282	3.492
d_2* for p = 5	2.096	2.563	2.871	2.992	3.098	3.277	3.488

Next, add the Product Variance estimate and the Combined R&R Variance estimate to get the Total Variance estimate:

$$\hat{\sigma}_x^2 = \hat{\sigma}_p^2 + \hat{\sigma}_e^2$$

III. CHARACTERIZING RELATIVE UTILITY and INTERPRETING THE RESULTS

Compute the fraction of the Total Variance that is contributed by Product Variation:

$$\frac{\hat{\sigma}_p^2}{\hat{\sigma}_x^2}$$

Now you can characterize the ability of this measurement system to capture meaningful information <u>in this specific application</u> (specific device and characteristic being measured):

% of Total Variance that is contributed by Product Variation	% of Total Variance that is contributed by Combined R&R	"Monitor Class" (or rating)
>80%	<20%	First Class
>50% and up to 80%	20% or more, but <50%	Second Class
>20% and up to 50%	50% or more, but <80%	Third Class
Up to 20%	80% or more	Fourth Class

The higher the class, the better the system's ability to loyally convey true unit-to-unit variation, and the lower its tendency to introduce and amplify noise (variation actually coming from the measurement system itself). First Class Monitors should be considered good, and Second Class Monitors may generally be considered acceptable (with caution proportional to its actual performance figures). Lower classes should trigger improvement if possible, or looking for alternatives.

Is the data useful for statistical inference?

In his article ***The Secret Foundation of Statistical Inference***, Dr. Wheeler makes a strong argument for the importance of ensuring that the data is coming from a *Homogenous Process*, before proceeding to any statistical analysis[5] (again, the entire original article is highly recommended):

"While the techniques of statistical inference were developed under the assumption of homogeneity, they make no attempt to verify that assumption. The formulas used in statistical inference are almost always symmetric functions of the data. Symmetric functions treat the data without regard to the time order of those data. (A change in the order of the data will not change the value of a symmetric function of those data.) Symmetric functions effectively make a very strong assumption of homogeneity. As a result, any lack of homogeneity will undermine the interpretation of the results."

*"Virtually every statistical technique is developed using the assumption that, on some level, you are dealing with **independent and identically distributed*** [i.i.d.] *random variables. Because of this, the question of whether or not your data display the appropriate level of homogeneity has always been, and will always be, **the primary question of data analysis**.*

*This question trumps all other questions. It trumps questions about which probability model to use. It trumps questions about how to torture the data with transformations. It trumps questions about what alpha level to use. In truth, you cannot define an alpha level, you cannot fit a probability model, and you cannot hope that your statistical inferences will work as advertised **if you do not have a homogeneous set of data**."*

Dr. Wheeler also states in that article that the Process Behavior Chart is the premier technique for empirically checking for homogeneity, because it's skeptical about the homogeneity assumption and explicitly examines the data for evidence of a lack of homogeneity.

"Any analysis is seriously flawed when it does not begin with a consideration of whether or not the data display an appropriate degree of homogeneity."

Further, in another article of his – ***Rational Sampling*** – Dr. Wheeler highlights the use of the *Individual Values and Moving Range Chart* (XmR Chart) as an elegant solution for the data subgrouping question[6]. Since the subgroup size is minimized (size 1) no assumptions about subgroup homogeneity is necessary – a subgroup with a single datapoint is perfectly homogenous – and the sensitivity of the chart is maximized (as pointed out by Shewhart[7], who is popularly perceived as the founding father of process behavior charts). However, as Dr. Wheeler says, "*because we will still be using pseudo-subgroups of size two defined by the successive individual values, we aren't totally free from the requirements of rational subgrouping. We will still need for successive values to be logically comparable.*" For this reason, it's desirable to have samples that are sequentially manufactured (or close to that), and important to keep track of their manufacturing order. If more than one datapoint is utilized per device (e.g. when testing complex and expensive capital equipment prototypes), the datapoints collection order should also be maintained in the charting process.

For each operator:

1. List all results (from that operator) in order (n results from 1st part, n results from 2nd part, etc).
2. Calculate the moving range (mR) between each pair of consecutive results: $mR_i = |x_{i+1} - x_i|$
3. Calculate the average mR.
4. Calculate the Upper Range Limit: $URL = D_4|_{n=2} * Av(mR_i) = 3.267 * Av(mR_i)$
5. Draw the mR chart – plot the mR_i figures in order, and mark the CL (average mR) and URL.

6. Calculate the *Individuals* (individual results) average, $Av(x_i)$
7. Calculate the Upper Control Limit: $UCL = Av(x_i) + 3 * (Av(mR_i) / d_2|_{n=2}) = Av(x_i) + 3 * Av(mR_i) / 1.128 = Av(x_i) + 2.660 * Av(mR_i)$
8. Similarly, calculate the Lower Control Limit: $LCL = Av(x_i) - 2.660 * Av(mR_i)$
9. Draw the Individuals (X) chart – plot the individual results in order, and mark the CL (Individuals average), UCL and LCL.

[In deriving UCL and LCL in steps 7&8 above, the average mR is adjusted through division by d_2 for a sample size n=2 (each mR is taken as a subgroup of size 2), to make it an appropriate measure of dispersion (the adjusted average mR can be thought of as a substitute for the sample standard deviation); then multiplied by 3, in line with the familiar ±3 StdDev "natural process limits".]

The individual charts from all operators should look generally similar; if not, it's important to figure out why.

Points outside the control limits or runs of points on one side of the CL indicate that the results potentially come from a non-homogeneous (unpredictable) source. When this occurs it's necessary to investigate the cause(s) before proceeding. As a first step consider reordering the results from each operator: p results from the 1st day (1st round of measurements), p results from the 2nd round, and so on; and recreating the XmR Chart (steps 2-9). Comparing the two versions of XmR charts may (or may not) reveal something meaningful – they each highlight the measurement process and the production process (each of which may independently contribute to unpredictability) in different ways.

As explained by Dr. Wheeler in several articles, this method doesn't rely on assuming that the data is Normal.

***

Once it is established that the measurement system can be relied on, and that the data at hand is coming from a homogeneous (predictable) source, trying to make some statistical predictions begins to make sense.

How likely are device units to meet the requirements?

The following method is based on the concept of ***Statistical Tolerance Intervals***. But while the "traditional" Tolerance Intervals approach begins with determining the desired *Coverage* and *Confidence Level* and then proceeds to calculating the necessary sample size (often resulting in sample sizes impractical for Design Verification), the method suggested here begins with a given (small) sample size and data already obtained, and looks for conclusions about Coverage-Confidence Level pairs expected inside the specified Design Input limits. The result is a prediction of the likelihood of the design (actually, an expression of the design – the product) meeting a Design Input requirement. If acceptance limits are set upfront (in the Design Verification Plan or protocols) for Coverage-Confidence Level pairs, such predictions can be used for Pass/Fail determinations for specific Design Input requirements (recall "*ensure*" from ISO 13485:2016 s. 7.3.6).

The method below is derived from the contents of D. J. Wheeler's article ***Statistical Tolerance Intervals***[9]. Again, for completeness and a better understanding of the theoretical foundation it's highly recommended to read the entire article.

According to Dr. Wheeler, the topic of Tolerance Intervals *"…requires very complex computations involving numerical integrations. What follows is based on a book of tables by Robert E. Odeh and Donald B. Owen published in 1980."* (the latter is apparently reference [10] in my references list). Additional, fairly-current support / endorsement of the concept can be found in ISO 16269-6:2014[11]. According to

Dr. Wheeler, *"…the underlying mathematics has not changed* […] *since Odeh, Owen, and others worked out the mathematical foundations."*

The general formula for a Tolerance Interval is: **Sample Average ± k * (Sample StdDev)**

k:: a factor depending on the Confidence Level, the Coverage, and N:: number of datapoints used.

Once the sample's average and sample's StdDev are computed (for the N datapoints), the largest k value that doesn't cause the Tolerance Interval to exceed the Design Input requirement (or *Specification Limits* in ISO's current relevant terms) can be found:

k = Min {(USL – Sample Av),(Sample Av – LSL)} / (Sample StdDEv)

USL:: Upper Specification Limit
LSL:: Lower Specification Limit

(This is based on an assumption that the sample average is within the Specification Limits.)

Now the obtained k value can be compared with the table below (derived from Wheeler, based on Odeh & Owen):

	Confidence Level = 0.99			**Confidence Level = 0.95**			**Confidence Level = 0.90**		
N	P = 0.90	P = 0.95	P = 0.98	P = 0.90	P = 0.95	P = 0.98	P = 0.90	P = 0.95	P = 0.98
5	6.655	7.870	11.10	4.291	5.077	7.165	3.499	4.142	5.851
10	3.617	4.294	6.102	2.856	3.393	4.827	2.546	3.026	4.306
15	2.967	3.529	5.030	2.492	2.965	4.230	2.285	2.720	3.881
20	2.675	3.184	4.546	2.319	2.760	3.943	2.158	2.570	3.672
30	2.394	2.851	4.077	2.145	2.555	3.654	2.029	2.417	3.457

Under each of the Confidence Levels (0.99, 0.95 and 0.90), looking at the row corresponding with the number of datapoints at hand (N), the obtained k value should fall somewhere among the 3 listed figures. The highest figure on that line (under each Confidence Level), that is still smaller than the obtained k value indicates a minimal Coverage (P) that can be predicted at that Confidence Level. A given k figure may fit in multiple locations (under different Confidence Levels), and thus more than one statement might be possible.

For example, if N = 10 and k = 3.1 (obtained as described above), the following statements can be made:

From the ***Confidence Level = 0.99*** table:	"The percentage of the population that falls within [Sample Average ± 3.1 * (Sample StdDev)], and thus within the Specification Limits (the Design Input requirement), at a probability of **99%**, is **<90%**."
From the ***Confidence Level = 0.95*** table:	"There's a **95%** probability that at least **90%** of the population falls within [Sample Average ± 3.1 * (Sample StdDev)], and thus within the Specification Limits (the Design Input requirement)."
From the ***Confidence Level = 0.90*** table:	"There's a **90%** probability that at least **95%** of the population falls within [Sample Average ± 3.1 * (Sample StdDev)], and thus within the Specification Limits (the Design Input requirement)."

In this case, the 3rd statement may be the most useful one: "There's a 90% probability that at least 95% of the population falls within the Design Input requirement." This may or may not be sufficient as *"ensuring"* that the Design Output satisfies the Design Input – depending on the policy (acceptance criteria) defined **upfront**. Under different criteria, the 2nd statement above may be more relevant; it may be a matter of policy.

As explained by Dr. Wheeler, Odeh & Owen's tables were developed for data coming from a Normal distribution; however, under certain approximations (explained in detail in his article) the figures can be used in most cases with data from any continuous unimodal distribution[9]. The table above is adapted from Odeh & Owen's, based on Wheeler's arguments, and thus it allows drawing conclusions regardless of the distribution the data is coming from. Hence, matching the data at hand with any specific distribution, or assuming one, is not necessary for using this method. It's important to note, however, that moving upwards along the Coverage scale (0.90 to 0.95 to 0.98), the approximation element becomes more pronounced; nevertheless, the predictions are still statistically sound for practical intents and purposes, at all three Coverage levels noted.

Sample size note: If, for example, one operator measured 5 parts, twice each, 10 datapoints are available for analysis as above (N=10). If the previous analysis stages characterized the measurement system as a good one (e.g. a "First Class Monitor" or a "Second Class Monitor" with a relatively low percentage of Total Variance contributed by Combined R&R), and >1 operators participated, datapoints from multiple operators may be combined, to achieve a larger N (20 or 30). For a given k figure, a larger N may indicate a larger Coverage under a given Confidence Level.

In his article's summary (Figure 8), Dr. Wheeler provides another table which allows determining a 95% Confidence Level Tolerance Interval for many more sample sizes in the range 10-30, predicting the Coverage within the interval, provided that the process is Homogenous[9]. The limits of this Tolerance Interval are the ones already calculated for the *Individuals* chart (UCL & LCL):

$$\mathbf{Av(x_i) \pm 2.660 * Av(mR_i)}$$

The table below is a subset (of the table in Figure 8[9]), containing some N figures that may be relevant to the process described here:

Confidence Level = 0.95	
N	P (%)
10	88.7
12	91.3
14	92.8
15	93.4
16	93.9
18	94.8
20	95.4
21	95.6
30	97.0

For example, if 2 operators measured 4 parts (device units), 2 times each, a total of 16 datapoints are available. If the XmR Chart indicates that the 16 datapoints come from a Homogenous process, it can be concluded that there is a 95% probability the Individuals Chart UCL & LCL [$Av(x_i) \pm 2.660 * Av(mR_i)$] bracket approximately 94% of the entire population that the sample represents. If the same interval falls within the Design Input requirement, this may be sufficient as *"ensuring"* that the Design Output satisfies the Design Input – again, depending on acceptance criteria defined **upfront**.