Commission of Inquiry into Construction Works at and near Hung Hom Station Extension under the Shatin to Central Link Project ("Original Inquiry")

Expert Report
Prepared by Barrie Wells

13 September 2019

CONTENTS PAGE

1. Introduction

1.1. I have been engaged by O'Melveny & Myers on behalf of Leighton Contractors (Asia) Limited ("**Leighton**") to provide statistical expert evidence for the Original Inquiry in relation to the Final Report on Holistic Assessment Strategy for the Hung Hom Station Extension [OU5/3229] ("**Holistic Report**").

1.2. My relevant area of expertise is in quality assurance testing and the statistical analysis of sample based testing for the purposes of assessing conformance with standards.

1.3. I hold a B.Sc. in Mathematics and Statistics from the University of Bath, an M.Sc. from Cranfield Institute of Technology and a Ph.D. in Underground Stress Analysis from University of Nottingham, England.

1.4. I have worked in statistical analysis and modelling for numerous government agencies and commercial organisations in Australia, North America and Europe, and have acted as a visiting lecturer in geostatistics at Nottingham University. Since about 1993, I have managed an independent company in North Wales, which provides services including mathematical and statistical analysis to companies worldwide. I have also provided statistical advice to standards committees in the UK and internationally.

1.5. I have provided a copy of my CV to the Commission.

2. Scope of Instructions

2.1. I have been instructed to:

(a) evaluate the efficacy of the sample based quality assurance testing that has been conducted on coupler connections (e.g. the PAUT tests);

(b) draw appropriate conclusions based on the results of such testing; and

(c) comment on the statistical analysis in relation to coupler connections that is set out in the Holistic Report (including at Sections 3.3, 4.2.3 to 4.2.4, 4.4 and Appendix B3).

3. Executive Summary

3.1. The Holistic Report addresses:[1]

(a) the sampling strategies that were used to obtain data for the purposes of the coupler engagement investigation; and

(b) the use of that data to calculate a strength reduction factor that should be applied to the EWL and NSL slabs.

3.2. In my opinion, the sampling strategies that were used to obtain the data are biased towards higher numbers of defective coupler connections ("**defectives**") such that any results obtained from the sampling (i.e. the results in Appendix B3) will lead to a more conservative result for structural competence and a higher than necessary strength reduction factor.

3.3. It is also my opinion that the methodology adopted for calculating the strength reduction factor is flawed. I have presented an alternative methodology that is, in my

[1] The relevant sections of the Holistic Report are Sections 3.3, 4.2.3 to 4.2.4, 4.4 and Appendix B3, and paragraphs 8 to 10 of the Executive Summary.

opinion, based on sound statistical principles. I have also presented an analysis of the data based on similar principles to that in the Holistic Report in order to provide an easier basis for comparison.

3.4. In my opinion, and relying on the threshold engagement length specified by the structural engineering experts, the correct defective rates that should be deduced from the results set out in Appendix B3 should result in strength reduction factors of 14.5% (EWL), 6.5% (NSL) and 9.4% (combined EWL and NSL). In my opinion, the correct approach is to take the combined sample resulting in a strength reduction factor of 9.4%.

3.5. The proposed suitable measures for Area A (Section 4.4.3 and Appendix C5 of the Holistic Report) are based on data that has been analysed in similar ways. There are further flaws in the analysis due to the need to combine the results obtained independently on two sides of the couplers. In my opinion, the strength reduction factor that should be applied to the coupler connections between the capping beam and EWL slab at Area A is at most 46.7% (as compared to 68.29% adopted by the Holistic Report) and is likely lower than that if the data was made available for the purpose of correcting the calculations.

3.6. The adoption of 95% as the confidence threshold has not been discussed in the Holistic Report. It is my opinion that, given the amount of difference this makes to the calculated values on which decisions are posited, some consideration should have been applied to this decision.

3.7. My comments on the sampling strategies and methodology used to calculate the strength reduction factor are presented in section 4 below under six main headings:

(a) Point 1: Sampling prior to testing couplers;

(b) Point 2: Acceptance and rejection: (a) PAUT test;

(c) Point 3: Acceptance and rejection: (b) Reasons for rejection;

(d) Point 4: Acceptance and rejection: (c) Rejection criterion;

(e) Point 5: Defective rate and Strength Reduction to be applied to connections between the capping beam and the EWL Slab at Area A; and

(f) Point 6: Consideration of the Appropriate Confidence Level

4. Analysis

Point 1: Sampling prior to testing couplers

4.1. Section 3.3.1 of the Holistic Report states:

> *"With reference to Clause 6.4.2 in the Holistic Proposal, the extent of Purpose (ii) opening-up works was based on a statistical approach with random sampling to assess the workmanship of the steel bar and coupler connections between the platform slabs and D-walls. A sample size of not less than 84 randomly selected couplers each for the EWL and NSL slabs would give a result with 95% confidence level using binomial statistics. To this end and with reference to Table 6.2 and Clause 6.4.24 in the Holistic Proposal, 28 locations each for the EWL and NSL slabs were randomly selected and a total of 56 selected locations were therefore to be opened up. It is expected that with at*

least three couplers at each selected location, not less than 84 randomly selected couplers at each slab, totalling at least 168 couplers, are needed to be inspected to ascertain the workmanship of the steel bar/coupler connections."

4.2. Sampling is a difficult subject to get right and is one of the more subjective aspects of mathematical statistics. However, *a posteriori* analysis is purely mathematical and we can ask the question: What is the probability that the sampling method chosen for assessing coupler competence was indeed random?

4.3. The document titled 'D-walls/Platform Slab Connections via Capping Beams' ("**Capping Beam Document**") [OU7/9805-9810][2] states that:

> "*From the construction record (out of total 237 D-wall panel), 175 nos. of them are without capping beams details (Type a) and 62 nos. of them are with capping beam details (Type b).*
>
> *...*
>
> *By the total number of samples in EWL, the random sample size (n) is 90 as obtained from the investigation, the number of Type a samples is 83 (na) and the number of Type b samples is 7 (nb) as from the randomly selected locations."*

4.4. In other words, the construction record shows a total of 237 D-wall panel connections, of which 175 are without capping beams details (Type A) and 62 are with capping beam details (Type B). A random sample of size 90 found 83 Type A specimens and 7 Type B specimens (these figures are taken from the Capping Beam Document).

4.5. The probability that this sample was random can be estimated using a hypothesis test for a population proportion (under the normal approximation assumption[3]). This test shows that the probability of it being a genuinely "random" sample is less than 1 in 1000. It is never the case that we can make definitive statements that a sample is or is not random; we can only say what the probability is that such a statement is true or false. It is therefore useful to have a frame of reference, to be able to compare this probability (of happening less than 1 time in 1,000) with other probabilities. By way of comparison, it is 100 times less likely than the thresholds set by the SC2:1995 (see e.g. Table 8 header) and SC2:2012 (see e.g. 1.6.1 or Table 9 footer) as the probabilistic threshold for acceptance of test results. We can therefore say that it is highly unlikely that this sample is random.

4.6. This result has ramifications, particularly for the calculation of the sample mean and sample variance used as estimators of the population from the small sample size for Type b specimens. In particular, the binary method, as chosen for the analysis in the Holistic Report, is more susceptible to lack of randomness than a method based on continuity, wherein a sample that fails the test by only a small amount is not completely discarded but is assumed to contribute to a reduced degree.

4.7. Another important question to ask is: Are the samples truly independent? The statistical technique used (i.e. binomial distribution) assumes independence. If, as is

[2] I am instructed that the 'D-walls/Platform Slab Connections via Capping Beams' document was prepared by MTRCL and relates to Section 3.3.19 of the Holistic Report.

[3] The normal approximation assumption is a means to simplify calculations on Binomial Distributions. See Note 5 for further details.

stated in Section 3.3.27 of the Holistic Report, a major reason for defects is poor workmanship, then defectives will probably be in clusters, and therefore not independent. This will lead to higher rates of defectives in the sample than in the population and hence any results (e.g. of strength reduction factors) will necessarily be more conservative than should be the case.

Point 2: Acceptance and rejection: (a) PAUT test

4.8. The note at the end of Table B3.1 in Appendix B3 of the Verification Report states:

> *"# The direct measurement result will take precedence over the PAUT result and as such it is considered as "defective" if the engagement length by direct measurement is less than 40mm"*

4.9. We should ask "How many times were the discrepancies significant?" There is insufficient information given in the Holistic Report but, if this is relevant, then it should be noted that, whilst a 3mm tolerance was assumed for PAUT measurements based on the accuracy of the method, direct measurements were apparently assumed to be infinitely accurate. It is noted that 36% of direct measurements in the EWL data set and 100% of direct measurements in the NSL data set are inconsistent with the inferred measurement back-calculated from the recorded number of threads exposed (after allowing for either the upper or lower limit to be used for comparison). That is, the direct measurements are inconsistent with both ends of the recorded range of number of threads, based on 4mm per thread and 44mm thread length. This is a strong evidential indication that the direct measurements should be assigned a tolerance for their use in a categorisation of 'defective' or 'not defective'.

Point 3: Acceptance and rejection: (b) Reasons for rejection

4.10. A number of test results in Appendix B3 of the Holistic Report have been discarded: see Item No. 1 to 12 of Table B3.1 and Item No. 1 to 6 of Table B3.2. The remarks for these test results are either: "*No valid PAUT result obtained, direct measurement cannot be obtained as the coupler cannot be unscrewed & sample / result discarded*" or *"No valid PAUT result obtained & sample / result discarded"*.

4.11. There are two different reasons given for the rejection of test results (i.e. for classifying a test as a 'failure') in Appendix B3 of the Holistic Report. These are either: (i) visual inspection of the connection between a rebar and a coupler; or (ii) measurement of the length of thread on a rebar that is embedded in a coupler. Measurements were only taken if visual inspection is passed. It follows that discarding a specimen because a measurement cannot be taken, when it has already passed the visual inspection (as, otherwise, no measurement would be attempted), biases the sample (see Note 1).[4]

4.12. The correct approach would be to adopt a Missing Values approach i.e. instead of discarding, treat them as missing and replace them with a representative or expected value (i.e. the statistically most likely outcome, had the measurement been successful). In Note 2, I explain why this is correct and give an example to show how doing otherwise biases the results of subsequent calculations.

4.13. Table 1 below shows the results if the 'missing value' approach is adopted.

[4] "Notes" are set out at the end of this report.

Point 4: Acceptance and rejection: (c) Rejection criterion

4.14. In this section, I will comment on the rejection criterion applied to the test results in Appendix B3 of the Holistic Report. My comments relate to:

(a) the use of a discrete rather than a continuous method;

(b) the 'number of threads exposed' criterion;

(c) the 'engagement length' criterion;

(d) an analysis of Appendix B3 when adopting a 28mm engagement length;

(e) an analysis of Appendix B3 when applying one criterion; and

(f) sample size for 95% confidence.

4.15. Section 3.3.13 of the Holistic Report explains the acceptance/rejection criteria as follows:

> *"For the purposes of this study, the proper couplers are considered to be*
>
> *(i) there shall be a maximum of two full threads exposed (which is stated in the manufacturer's installation requirements); and*
>
> *(ii) the engagement length of the threaded steel rebar inside the coupler should be at least 40mm. As the allowable measurement tolerance of the test equipment is 3mm, equipment readings below 37mm are regarded as defective."*

4.16. I will refer to the first of this criteria as the 'number of threads exposed' criterion and the second as the 'engagement length' criterion.

4.17. I understand that:

(a) the threaded length of Type A rebar would be typically 44mm and for Type B rebar would be typically 88mm;[5]

(b) the recorded threaded lengths of many rebar in Appendix B3 is less than 44mm. The percentage of rebar whose recorded thread length, made up of the combined length of the exposed threads and engaged thread, is less than 44mm is 63%, if taking the average exposed thread length, or up to 83% if looking at extrema. The variation arises because the exposed thread length is recorded as a range;

(c) one full thread is 4mm;[6]

(d) according to the manufacturer's installation requirements, it is acceptable for a maximum of two threads (i.e. 8mm) to be exposed on visual inspection;[7]

(e) if two threads are exposed on a rebar with a 44mm threaded length, the maximum engagement length of a typical rebar would be 36mm;

(f) the tolerance for the PAUT tests is plus or minus 3mm. Therefore, it is possible that a rebar with two threads exposed and a recorded threaded length of 33mm could still be engaged to the maximum extent possible. For example, the item no. 56 in Table B3.2 of the Appendix B3 has an engagement length of 33.1mm

[5] Section 3.3.18 of the Holistic Report.
[6] See Section 3.3.18 of the Holistic Report and Section 15.4 of the Expert Report of Mr. Nick Southward dated January 2019.
[7] See Section 15.4 of the Expert Report of Mr. Nick Southward dated January 2019.

and 1-2 threads exposed. It follows that it is possible that items in Appendix B3 which are rejected because they only satisfy one of the criterion noted above could actually be "not defective" and hence would provide their full design contribution to the structure.

4.18. In light of these facts, it is my opinion that the coupler connections referred to in Appendix B3 should be graded as "not defective" if they satisfy the 'number of threads exposed' criterion. These facts also cast doubt on whether the 'engagement length' criterion should be set at 37mm. For this reason, I have conducted an analysis in Table 2 below to determine the defective rates if either of these criteria is adopted.

Continuous or Discrete?

4.19. The engagement length when measured by the PAUT method has been assigned a 3mm tolerance. However, the "number of full threads exposed" has no tolerance (although it is quoted as a range). Furthermore, the number of threads exposed depends on the angle of observation. The relationship between the number of threads exposed and the effect on the structural integrity could reasonably be used to distinguish discrete and continuous metrics (see Note 7), wherein an appropriate choice would remove the ambiguity arising from the failure condition of "3 - 4 threads exposed" overlapping with the pass condition of "2 - 3 threads exposed".

4.20. In paragraph 3.6 of my report in the Extended Inquiry, I noted the dangers of applying discrete-value methods to continuous variables. Here, we have the situation that a coupler connection which does not meet a continuous criterion is subject to a binary pass/fail decision, whereas a small discrepancy in the measurement of engagement length or threads exposed would presumably result in a small reduction in contribution to the competence of the structure. Whilst such an assessment is not technically within the purview of mathematical statistics, the choice of using discrete statistical methods to describe a continuous situation is a statistical matter. In particular, it should be noted that errors in measurement or classification in continuous methods do not lead to bias, whereas errors in binary methods do lead to bias in the results obtained (see Note 3).

'Number of threads exposed' criterion

4.21. It should be noted that the failure condition of "3 - 4 threads exposed" overlaps the pass condition of "2 - 3 threads exposed", so that the same condition could be deemed a 'Pass' by one person and a 'Fail' by a different person.

4.22. It should be noted that in 20.3% of specimens reported in Table B3 (25.3% EWL, 15.2% NSL), the measurement of engagement length is not consistent with the recorded number of threads exposed, based on an assumed thread length of 44mm and length of a single thread of 4mm. I have therefore conducted an analysis (as set out in Table 2 below) of the defective rate if these inconsistent specimens are discounted. This is explained more fully in paragraph 4.33 below.

'Engagement length' criterion

4.23. Based on the test results in Table B3.1 and B3.2, the (robust) mean of the "Enhanced PAUT Engagement Length (mm)" of defective specimens is 34.8mm (for purpose (ii) at EWL) and 34.4mm (for purpose (ii) at NSL). This engagement length is only 5.9% (EWL) and 7.0% (NSL) less than the criterion of 37mm as adopted in the Holistic Report.

4.24. A rebar coupling with 37mm engagement length is assumed to be fully functioning, carrying the design load. However, a rebar coupling with 34.8mm engagement length (which is the mean for the EWL and only 5.8% less than the engagement length criterion) is assumed to bear no load and to be completely ineffective.

4.25. Taking these considerations into account, it can be shown that the strength reduction factors adopted in the Holistic Report of 36.6% (EWL) and 33.2% (NSL), which are based on the number of 'defectives', should be replaced by strength reduction factors of 9.1% (EWL) and 3% (NSL) (see Note 6 for detailed calculation). This should be further reduced by combining the EWL and NSL sample, which results in a strength reduction factor of 6.6% (see Note 6 for detailed calculation). I consider this figure of 6.6% to be a superior estimate to that stated in the Holistic Report.

4.26. I have also conducted an analysis of the test results in Appendix B3 based on similar principles to that in the Holistic Report in order to provide an easier basis for comparison. I have presented this analysis in the following sections of my report.

Analysis of Appendix B3: Adopting 28mm engagement length

4.27. I have conducted an analysis of the results in Appendix B3 by adopting an engagement length of 28mm (rather than 37mm). This analysis is presented in Table 1 below.

4.28. The expert evidence from the structural engineers[8] indicates that the threshold engagement length for structural integrity should be no more than 28mm. If this threshold is adopted instead of 37mm, the defective rates of coupler connections reduce significantly from those stated in the Holistic Report.

Table 1. Analysis of Appendix B3: Adopting 28mm engagement length

	Strength reduction factor in Holistic Report[9]	Assuming missing values have mean of the remainder of the sample	Adopting engagement length cut-off of 28mm	Engagement length cut-off 28mm and assigning mean to missing values
Table B3.1 Coupler Engagement Length Result for Purpose (ii) at EWL	0.366	0.339	0.163	0.145
Table B3.2 Coupler Engagement Length Result for Purpose (ii) at NSL	0.332	0.327	0.069	0.065
Tables B3.1, B3.2 combined	0.350	0.308	0.102	0.094

4.29. The first column in the table above reflects the approach set out in the Holistic Report and results in a defective rate of 36.6% (EWL only), 33.2% (NSL only) and 35%

[8] The minimum engagement lengths specified by these experts are: (i) 22mm to 24mm (Prof McQuillan's report at §119); (ii) 26.4mm (Southward's report at §15.5 and Glover's report at §7.2); and (iii) 28mm (Prof Yeung's oral evidence in which he accepted that at most 7 threads would be sufficient (Day 41:161(2)-(5)).

[9] The Verification Report adopts a strength reduction factor of 35%. As stated in my report for the Extended Inquiry, it appears that the figure of 35% is based on the combined samples of the EWL and NSL.

(combined EWL and NSL). The second column applies the same approach but factors in the samples that were discarded in the Holistic Report (by assigning them the "mean" value of the remainder of the sample) and results in a defective rate of 33.9% (EWL only), 32.7% (NSL only) and 30.8% (combined EWL and NSL). The third column adopts the same approach as the Holistic Report (i.e. the first column) but applies a threshold 'engagement length' criterion of 28mm and results in a defective rate of 16.3% (EWL only), 6.9% (NSL only) and 10.2% (combined EWL and NSL). The fourth column adopts the approach in the second column but applies a threshold 'engagement length' criterion of 28mm and results in a defective rate of 14.5% (EWL only), 6.5% (NSL only) and 9.4% (combined EWL and NSL).

4.30. In my opinion, and in reliance on the threshold engagement length specified by the structural engineering experts, the correct defective rate that should be drawn from the results set out in Appendix B3 should be as noted in the fourth column of the table above. It follows that the defective rates of 36.6% (EWL only), 33.2% (NSL only) and 35% (combined EWL and NSL) noted in the Holistic Report should reduce to 14.5% (EWL), 6.5% (NSL) and 9.4% (combined EWL and NSL).

4.31. In my opinion, the correct approach is to take the combined sample of EWL and NSL slabs, resulting in a defective rate of 9.4%.

4.32. It is noted that the acceptance criterion is steadily reduced. This not to be read as implying a slackening of the criterion but of tightening the rigour of the statistical method, which happens to result in a lowering of the threshold.

Analysis of Appendix B3: Applying one criterion

4.33. I have conducted a separate analysis of the results in Appendix B3 by adopting either one of the two criterion used in the Holistic Report, namely: (i) the 'engagement length' criterion (at least 37mm); or (ii) the 'number of threads exposed' criterion (maximum of 2 full threads exposed). This analysis is presented in Table 2a below. The details of my calculations are in Note 8.

4.34. In Table 2a below, the first column (as used in the Holistic Report) is based on a coupler connection being judged as defective if it fails either the 'engagement length' criterion (at least 37mm) or the 'number of threads exposed' criterion (no more than 2). From the data in Tables 2b and 2c below, it is apparent that this is not a good way to set the criteria because it is internally inconsistent and produces contradictory results.

Table 2a. Analysis of Appendix B3: Applying each criterion separately

	Strength reduction factor adopted by the Holistic Report	Strength reduction factor based on *'engagement length cut-off 37mm'* alone	Strength reduction factor based on *'no. of threads exposed ≤ 2'* alone	Strength reduction factor if both criteria are needed for a 'fail'
EWL	0.366	0.354	0.230	0.240
NSL	0.332	0.331	0.059	0.082
Combined	0.350	0.317	0.135	0.146

Table 2b. Analysis of Appendix B3: Frequency of Contradictions

	Percentage of specimens where the two measures contradict
EWL	25.3
NSL	15.2
Combined	20.3

Table 2c. Analysis of Appendix B3: Comparison of Measurement Methods

EWL, NSL Combined	Average	Extreme Range
% whose PAUT + Threads exposed = 44	2.3	
% whose PAUT + Threads exposed < 44	63.2	83.0
% whose PAUT + Threads exposed > 44	34.5	17.0

Sample size for 95% confidence

4.35. As noted above, Section 3.3.1 of the Verification Report states that "*A sample size of not less than 84 randomly selected couplers each for the EWL and NSL slabs would give a result with 95% confidence level using binomial statistics*."

4.36. Calculation of the number of specimens needed to yield a specified confidence (e.g. a 95% level of confidence) requires knowledge of the population size and variance or an estimate based on a known sample. Here, the variance is not known in advance and the analyst who designed the testing of the coupler connections therefore chose a worst case scenario (the least confidence arises if there is a 50% success rate, and hence also a 50% failure rate) and determined the number of samples that would be required in the worst case scenario. This means that 84 is an upper limit for the number required in order to provide an estimate with a 95% confidence level. In nearly all cases, the confidence would actually be higher than 95%. In fact, the actual confidence will be higher in every case except exactly equal numbers of successes and failures. Such an approach (using a worst case scenario) would be adopted if the only consideration was time (or effort) in sampling. Here, however, the sampling technique is destructive of the as-constructed structure and a more accurate methodology might have been better selected if safety were the primary goal (see Note 4).

4.37. In simple terms, this means that the number of coupler connections (84) that were tested as part of the sample was the maximum that needed to be tested in order to deliver useable results. If this study had been designed more accurately, fewer coupler connections could have been tested.

Point 5: Defective Rate and Strength Reduction to be applied to the EWL Slab at Area A ('connections between the capping beam and the EWL Slab at Area A')

4.38. The proposed suitable measures for Area A (Holistic Report section 4.4.3 and Appendix C5) are based on statistics not presented in the Holistic Report but made available by MTRCL in the Capping Beam Document (see paragraph 4.3 above). The statistical analysis presented in the Capping Beam Document concludes that a 68.29% strength reduction (with 95% confidence level) should be applied to take account of the condition on both sides of the couplers.

4.39. As with Table B3.1 and B3.2, the sample was biased by discarding samples that passed the first part of the test for determining whether defective (i.e. the visual inspection) but for which the second part of the test (i.e. measuring the length of the engaged thread) could not then be completed. In Table 3 below, I have set out the results that follow if these discarded samples are included as 'missing values' (see Note 3) and are analysed in same way as in the MTRCL's Capping Beam Document.

Table 3: Strength Reduction Factor, connections between the capping beam and the EWL Slab at Area A

			95% upper bound		
Defective:	Mean	Variance	Type A	Type B	Combined
MTR calculation	0.313	0.0031	0.358	0.683	0.405
Missing Values	0.295	0.0026	0.358	0.574	0.380

4.40. This shows the results of making one correction to the data that was input to the calculation used. There are other reasons for questioning the data input to this calculation, and hence other ways to re-calculate this strength reduction factor, but insufficient detail is given to allow me to conduct this re-calculation exercise. Other corrections may be required because:

(a) The criterion for determining whether a coupler connection is deemed defective was, apparently, to use the 'number of exposed threads' criterion on the Capping Beam Side and the 'engagement length' criterion (at least 37mm, based on the "Enhanced PAUT Engagement Length") on the EWL Slab Side. As noted above in paragraph 4.22, these two are frequently (20% of the time overall, rising to 25% in the area where this analysis was conducted) mutually incompatible. Therefore, to mix them in a single calculation, using each as a sole representative criterion, seems to me to be inappropriate, although there is so little detail given about this calculation and its input data that it is not clear whether this criticism is valid here;

(b) It is not clear to me why data relating to the EWL Slab Side does not also use the main EWL data set, as doing so would greatly increase the confidence in the results as well as overcoming some of the mistakes made (by assuming a large sample approximation, when the sample size was actually very small); and

(c) As with my comments on the coupler connections generally (see paragraphs 4.19 and 4.20 above), I am of the opinion that binomial analysis is not the best way to treat these continuous data sets for the purposes of analysing the connections in the EWL Slab at Area A.

4.41. In addition to the above criticisms of the data, the calculation based on the data is incorrect because:

(a) The analysis used in deriving these upper bounds assumes that the combination of two binomial distributions provides an additive relationship for the variances. This is true for chi-squared variables but not for binomial distributions unless the probabilities are identical (in which case the addition would not be needed); and

(b) The analysis assumes that the Delta Method approximation is valid, which it is not for this small sample (see Note 5 and paragraph 5.2 below).

4.42. In Table 4 below, I have set out the relevant figures after re-calculating with using the Monte-Carlo method and treating the discarded values as 'missing values'.

Table 4. Strength reduction factor to be applied to the EWL Slab at Area A

P_b (i.e. type b probabilities)	Mean	Standard Error	95% confidence Upper Bound
MTRCL (Capping Beam Document)	0.416	0.0264	0.683
Monte Carlo (see Note 8)	0.365	0.0615	0.467

4.43. It follows that the strength reduction factor to be applied to the EWL Slab at Area A should be lower than that which is adopted by the Holistic Report (and detailed in the Capping Beam Document). The proposed suitable measures in Area A (4.4.3 and Appendix C of the Holistic Report) should be re-assessed in light of this analysis.

Point 6: Consideration of the Appropriate Confidence Level

4.44. The calculations of the strength reduction factor in the Holistic Report follow the pattern of calculating the mean and standard deviation of a sample and then using these figures to derive a 95% confidence value (i.e. a value which we would not expect to be exceeded 95% of the time). Once a confidence interval has been chosen, the calculations are purely mathematical. There is no room for interpretation and anyone performing the calculation should arrive at the same answer. The confidence level or threshold to use is, however, generally subjective, and to be determined on external grounds. These grounds may include the degree of caution required, or the desire to minimise intrusive works.

4.45. The confidence level chosen does not give a measure of the safety factor; that is a matter for the engineers to determine, based on engineering principles. It gives us a level of confidence that the correct decision has been made. As with engagement length, it is not binary: a slightly wrong decision means a slightly greater chance of the wrong outcome.

4.46. Given the amount of subjectivity inherent in this choice, I consider the lack of discussion (on why the value of 95% was chosen) to be an omission. For guidance, we could look at CS2:1995 or CS2:2012, where the latter should now be the reference of choice.

4.47. CS2:2012 relevantly states:

> *"The characteristic values as given in Table 5 are (unless otherwise indicated) the lower or upper limit of the statistical tolerance interval at which there is a 90% probability (1-α = 0.90) that 95% (p = 0.95) or 90% (p = 0.90) of the values are at or above the lower limit or at or below the upper limit respectively. This quality level refers to the long-term quality level of production."*

4.48. Whilst it is by no means unequivocal, there is an indication here that the Standing Committee on Concrete Technology of the Government of the Hong Kong Special

Administrative Region would provide guidance that a confidence level of 90% should be adopted as appropriate for the statistical tolerance interval.

4.49. If recalculated with a 90% confidence level, the strength reduction factors in Tables 1 and 4 above reduce significantly. These figures are set out below in in Tables 5 and 6 respectively. I present these figures without further comment, primarily as an illustration of the effect on the results of this decision (on confidence level to choose), which was made externally to the Holistic Report.

Table 5. Analysis of Appendix B3: Adopting 28mm engagement length and 90% Confidence Level

	Strength reduction factor in Holistic Report[10]	Assuming missing values have mean of the remainder of the sample	Adopting engagement length cut-off of 28mm	Engagement length cut-off 28mm and assigning mean to missing values
Table B3.1 Coupler Engagement Length Result for Purpose (ii) at EWL	0.338	0.339	0.140	0.125
Table B3.2 Coupler Engagement Length Result for Purpose (ii) at NSL	0.305	0.327	0.056	0.053
Tables B3.1, B3.2 combined	0.304	0.308	0.089	0.082

Table 6: Strength Reduction Factor at connections between the capping beam and the EWL Slab at Area A

			90% upper bound		
Defective:	Mean	Variance	Type A	Type B	Combined
MTR calculation	0.313	0.0031	0.340	0.624	0.385
Missing Values	0.295	0.0026	0.340	0.524	0.361

5. Conclusion

5.1. The Executive Summary of the Holistic Report states (at paragraph 10):

> *"A total of 102 and 99 samples at EWL and NSL slabs respectively have eventually been examined. Among these, 90 and 93 samples at the EWL and NSL slabs respectively yielded valid results for statistical analysis. For the purpose of Stage 2b, engagement lengths found to be less than 37 mm by PAUT or 40 mm by direct measurement are treated as not complying with the manufacturer's installation requirements and are considered as defective. 25 out of 90 samples at the EWL slab and 23 out of 93 samples at the NSL slab*

[10] The Verification Report adopts a strength reduction factor of 35%. As stated in my report for the Extended Inquiry, it appears that the figure of 35% is based on the combined samples of the EWL and NSL.

were defective. Based on the binomial analysis, it is estimated that, with a 95% confidence level, not more than 36.6% and 33.2% of couplers at the EWL and NSL slabs respectively are considered defective."

5.2. In my opinion, binomial analysis is not appropriate to the data domain, the 95% confidence limit has been calculated inappropriately and the acceptance and rejection criteria are not valid. If the binomial analysis were corrected and improved, the calculated defective rate of coupler connections would actually be about one quarter of that calculated in the report (0.09 instead of 0.36). The recommendations in sections 4.3.6 and 4.3.7 of the Holistic Report should therefore be re-assessed in the light of those recalculated results.

5.3. In my opinion, the correct way to conduct this analysis would be on the strengths predicted from the measurements taken on the specimens in the sample, to enable the structural engineers to make a direct assessment. This data is not available but, if it were, it would be expected to provide a lower estimate for a Strength Reduction Factor and hence the result obtained, of a strength reduction factor of 9.4%, should be treated as an upper bound (or, in terms of structural safety, a very conservative estimate).

5.4. In my opinion, the sampling method adopted was: (i) non-optimal (in the sense that a larger sample was taken than was necessary for the stated confidence objectives); and (ii) biased towards a higher number of defectives than should be expected in the population (i.e. the as-constructed structure). For both reasons, there should be no need for further sampling since both over-sampling and bias work to increase the confidence in the safety margin created.

6. Expert Declaration

6.1. I understand that my primary duty in preparing this report and giving evidence is to the Commission of Inquiry, rather than to the party who engaged me and I have complied with that duty.

6.2. I have endeavoured in this report and in my opinions to be accurate and to have covered all relevant issues concerning the matters stated which I have been asked to address.

6.3. I have endeavoured to include in my report those matters, which I have knowledge of or of which I have been made aware, that might adversely affect the validity of my opinion.

6.4. I have indicated the sources of all information that I have used.

6.5. I have not, without forming an independent view, included or excluded anything which has been suggested to me by others (in particular my instructing solicitors).

6.6. I understand that:

(a) My report, subject to any corrections before swearing as to its correctness, will form the evidence to be given under oath or affirmation.

(b) I may be cross examined on my report by a cross examiner assisted by an expert.

(c) I am likely to be the subject of public adverse criticism if the COI concludes that I have not taken reasonable care in trying to meet the standards set out above.

6.7. I believe the facts I have stated in this report are true and that the opinions I have expressed are correct.

Barrie Wells

Digitally signed by Barrie Wells
DN: cn=Barrie Wells, o=Conwy Valley Systems Limited, ou=Cowny Valley Systems Limited, email=Barrie.Wells@ConwyValley.com, c=GB
Date: 2019.09.13 09:49:44 +01'00'

BARRIE WELLS

13 SEPTEMBER 2019

Note 1: Terminology

"Independence"

In order to assess the probability of an unknown event happening, we generally can only rely on historical information. If the same thing has been observed to have happened 3 times previously, out of 12 occasions on which it could have happened, then the probability of it happening in the future is best estimated as 3 out of 12, or a 25% possibility.

When considering the probability of an unknown event happening, we need to know whether each event is independent, or whether each one has an effect on the probability of the next. There is a common misconception that, if a coin comes down 'Heads' 10 times in a row, then it is more likely to be tails next time, because overall it needs to have the same number of each so it has a lot of catching up to do, whereas in fact (as the coin has no memory), each event is independent of the previous.

"Bias"

Continuing the example above: in practice, the opposite is more likely to be true: 10 Heads in a row is good evidence that the coin is biased and hence is more likely to continue to show Heads. In practical situations, wherein no theoretical probability can be calculated *a priori* (unlike the case of tossing a coin), bias is difficult, if not impossible, to detect purely from analysis of results. It is therefore important to be critical of the method, and in particular the sampling technique, in order to determine potential sources of bias.

As an example, telephone surveys are intrinsically biased, because they only sample the subset of the population with a telephone. Further, a company conducting a survey may employ people during the day to make the calls, adding to the bias by only sampling people who both can and will answer the telephone during the day. Thus, answers to the survey question are only provided by people who have previously satisfied a condition that may or may not be relevant to the survey; if someone both can and will answer the telephone during the day, then answers to questions on their place of work will likely not be representative of the population as a whole, they will be biased, and any decisions made on the survey results could prove disastrous for the company using them.

This situation ('both can and will answer the telephone during the day') is analogous to the testing of rebar connection engagement lengths: engagement length tests are only attempted on connected couplers, so the set of discarded specimens is not a subset of the entire population (or even of the sample), it has already been filtered. This situation, and how to handle it, is discussed further in Note 2.

Note 2: Handling Missing Values

If some specimens do not yield a value, they can either be discarded or be replaced with the mean of the rest of the sample. Both approaches should result in the same measure of estimated mean, although the latter will lower the estimate of the variance. In some situations, either may be chosen without consequence.

A decision on how to treat specimens that do not yield a value, either discarding completely or treating as a missing value, becomes critical when those specimens are not randomly distributed in the population. In such situations, it is necessary to account for the reasons why there is no answer.

Take for instance a telephone survey asking 100 people, first whether they are male or female, and then asking if they have a beard. 50% of respondents will be women, so they are assumed not to have a beard, without asking. Of the remaining 50%, i.e. the men, assume half say "Yes", so we calculate an estimate of 25% of the population have a beard. If, however, 40% of men decline to answer, but the remaining 60% (i.e. 30 men) are representative so again half of them say "Yes", then we calculate an estimate of less than 20% of the population have a beard (half of the 60% of men who answered is 15 men, plus the 50 respondents who are women, makes 65 out of the 80 who answered do not have a beard, and 15 out of 80 do; 15/80 = 18.75%).

The correct way to calculate would be to normalise. There are several ways to do this, all equivalent, so assume that the 20 men who did not answer would have answered in the same proportions as those who did; i.e. assign, to the 'missing' values, the typical or expected or average values calculated from the results we did obtain.

Note 3: Binary Methods

A binary method is a special case of a discrete method, in which only two possibilities can occur. Throwing a dice is an example of a discrete statistic, tossing a coin is an example of a binary statistic. 'Discrete' is used here in contrast to ‘Continuous’. In many situations, there is no alternative to using discrete methods because the data are non-numeric. In situations where a choice may reasonably be made to use either discrete or continuous, there are generally advantages to using continuous methods, including:

1. Discrete approximations to continuous situations lead to a loss of accuracy, as information is being ignored. This can be seen simply by looking at rounding real numbers to integers, leading to a loss of detail which may or may not be important but its importance should be considered and the basis for decisions documented.
2. Errors in measurement or classification in continuous methods do not lead to bias, whereas errors in binary methods do lead to bias in results obtained. (see e.g. 'A General Approach to Analysing Epidemiological Data That Contain Misclassification Errors", Espeland, M.A, and Hui, S. L., Biometrics 43, 1987)

Note 4: Estimating Sample Size When Variance is Unknown

If variance is not known in advance, and cannot be estimated, one method for optimising the number of specimens needed to form a sample with a required confidence is to adopt a 'stopping criterion'. This involves starting the sampling with an estimated sample size based on a best estimate (or best guess) of the variance, then continuously monitoring the variance after each measurement has been taken and adjusting the best guess accordingly, so that a guess becomes successively more accurate, and approaches a proper statistical estimate.

Such an approach is important if the cost of sampling is high either in terms of time and/or effort or in the damage caused by sampling.

Note 5: The Normal Approximation to the Binomial Distribution

The normal approximation assumption is a means to simplify calculations on Binomial Distributions. It is used because of the numerical instability in direct Binomial calculations, which depend on raising the probability to a power up to the size of the number of specimens. The approximate calculation replacing the power formula depends on ignoring higher order terms in a Taylor's Series expansion, so is only valid if the ignored terms are very small in comparison with the included terms. A rule of thumb is that the approximation is valid if the sample size (or population size, as it is applied either to samples or populations) multiplied by the probability is at least 10 and that the sample or population size multiplied by one minus the probability is also at least 10

5.1 In the example of 4.5, we see that the rule of thumb says that the approximation is safe to apply because the population size is 237 and the two probabilities (P and (1-P)) are 0.74 and 0.36.

5.2 In the case of 4.3.2 (Area A calculation), n =7, P = 0.74, so nP = ≈ 5

Note 6: Calculation of Strength Reduction Factor based on Continuity of Contribution

The Holistic Report calculates a strength reduction factor by assuming that an engagement length less than 37mm (the cut-off or threshold used in the Report) renders a coupled rebar unable to contribute to the strength of the structure. If we accept that a small reduction in engagement results in a corresponding reduction in contribution to strength, then we can calculate the mean engagement length, capped at 37mm (i.e. any coupler with an engagement length greater than 37 mm is assumed to have exactly 37mm) and use the ratio as an indicator of required strength reduction factor. This may not be the approach preferred by a structural engineer but it is in my opinion superior to that of the binomial approach in statistical terms, as being a more valid representation of the available data.

Replacing the 'Missing Values' with the mean of the remainder of their class (i.e. specimens for which measurements were successfully taken), the mean of all EWL data is 33.6 mm, resulting in a strength reduction actor of 9.1%. The mean of all NSL data is 35.5 mm, resulting in a strength reduction factor of 3%.

The mean of all engagement length data, treating uncoupled or cut rebar as zero and missing values as the mean of all measured engagement lengths, combining EWL and NSL, is 34.6 mm, resulting in a strength reduction factor of 6.6%.

Note 7: Continuous and Discrete Variables

The distinction between continuous and discrete variables is important because of the decision in the Holistic Report to treat rebar couplers as either defective or not defective, i.e. a binary choice, similar to 'Heads' or 'Tails' when flipping a coin. In the case of a coin, there is no other choice, unless we allow for the coin landing on its edge, in which case it is still discrete but now ternary. Although the flipping of a coin is the most frequently used example of a binary variable, there is no reason why a binary choice variable should be equally likely to occur in either outcome. In the Holistic Report, the probability of 'defective' outcomes in the population (i.e. the structure) is estimated by dividing the number of specimens examined by the number of defectives seen.

Engagement length is a continuous measurement and it seems surprising that the Holistic Report chose to treat it not only as discrete, but as binary. It is true that sometimes a continuous variable has to have a discreet cut-off applied. For example. An examination which results in either a Pass or a Fail. However, such situations are invariably fraught with difficulty and subject to frequent recounts. Such a decision is forced by the situation, not chosen as the best option. In the UK, breaking of speed limits incurs a fine based on the amount by which the limit was exceeded. There would be much unfairness if a small infraction were punished the same as a large infraction.

In structural situations, using such an approach could lead to a situation in which a larger number of small supports would not be allowed to replace a smaller number of large supports, severely limiting architectural possibilities.

Statistically speaking, a discreet variable is subject to greater constraints to ensure fairness (lack of bias). This can be proven theoretically (see references in Note 3) but may better be seen heuristically: in all cases except a perfectly uniform distribution, any boundary imposed on a continuum will necessarily have a bias towards whichever side of the distribution is higher at that cut-off point.

The number of threads exposed is put into a categorical variable in the Holistic Report: [0,1], [1,2], [2,3], … etc. rather than being treated as continuous. This may be seen as a more reasonable decision than the decision on engagement length, but here we must be aware that assessment is subjective, dependent on angle of view as well as eyesight.

In neither situation do I believe that the case for treating these variables as discrete has been made, statistically.

Note 8. Analysis of Appendix B3: Applying each criterion separately

		Holistic Report: Table B3 criterion	'No. threads exposed ≤ 2' criterion	Assign Missing Values
Section 1: Data				
EWL	Sample size	90	102	102
	no. discarded	12	0	0
	no. defective	25	17	25
	No. not defective	65	98	77
NSL	Sample size	93	99	99
	no. discarded	6	0	0
	no. defective	23	3	22
	No. not defective	70	96	
Combined	Sample size	183	201	201
	no. discarded	18	0	0
	no. defective	48	20	47
	No. not defective	135	194	77
Section 2: Analysis				
EWL	Mean (np)	0.278	0.167	0.245
	Variance {n.p.(1-p)}	0.0022	0.0016	0.0018
	95% confidence limit	0.355*	0.232*	0.315 *
NSL	Mean (np)	0.247	0.030	0.222
	Variance {n.p.(1-p)}	0.0020	0.0003	0.0000
	95% confidence limit	0.321*	0.059*	0.222 *
Combined	Mean (np)	0.262	0.100	0.234
	Variance {n.p.(1-p)}	0.0011	0.0005	0.0004
	95% confidence limit	0.316*	0.135*	0.269 *

* z-test, large sample approximation (see Note 5 for explanation of 'large sample approximation'.)

Note 9. Monte Carlo Analysis of Probabilistic Situations

Monte Carlo analysis is the preferred method for solving problems where the analytical method is intractable. For reference, standard textbooks such as Hammersley & Handscomb, (1964) may be consulted or, for a reference with an engineering basis, the Missouri University of Science and Technology Course Notes for Probabilistic Engineering Design:

http://web.mst.edu/~dux/repository/me360/me360.html

For a more statistical description of the requirement for its usage as employed here:

Chan, T. F., Golub, G. H., and LeVeque, R. J. (1983). Algorithms for computing the sample variance: Analysis and recommendations. The American Statistician, 37(3):242-247

Being a distribution-free method, we do not need to make assumptions on sample size.