The Importance of the Consistency Factor in CQT and Other Polygraph Tests[1]

Avital Ginton[2]

Abstract

Repeating questions and collecting several charts is routine in polygraph examinations. The paper asserts that repetitions is not just a “nice to have” element in the examination, rather it is a critical factor for decision making in CQT (and other polygraph techniques), if we want to ensure that the outcomes are not just a matter of chance. The logic behind repetitions is that they tend to nullify chance effects and leave the effects that bear consistency. Acknowledging the importance of a consistency factor which is manifested through the repetitions, a set of Between-Charts-Consistency Decision Rules is suggested which enables examiners to render a DI or NDI call based on the degree of consistency between charts with regard to the direction of the numerical outcomes (Plus, Minus or Zero). This approach which is conceptually different from the decision rules found in the common scoring methods such as the Federal Investigative or Evidentiary, or the Utah TDA models, was implemented on a sample of 1000 verified scorings from the NCCA database and reached a mean of decision accuracy of 87.81% (91.88% for truth-tellers and 83.40% for deceptive examinees), with an Inconclusive rate of 9.8%. The significance of repetitions is further discussed indicating that from the consistency factor perspective per-se, and in accordance to scientific convention ($\alpha = 0.05$), a three-chart test is not enough to avoid the possibility that the found outcome is a mere effect of chance fluctuations.

The premise behind psycho-physiological detection of deception (PDD) is that it is possible to differentiate between deception and truthfulness, or to be more precise, to detect the differential psychological states of mind of deceiving vs. telling the truth about certain facts, by measuring certain physiological activity that accompany them. This premise has led to development of various kinds of interrogation methods in which a suspect has been interrogated while being connected to a polygraph instrument. These techniques have eventually turned into more structured tests in which the interrogator asks the suspect several pre-planned questions in a preplanned manner and order. Along this development, the interrogator has become an "examiner" and the suspect, an "examinee." These changes in the language are not just cosmetic, rather they reflect the intention to move toward a more objective and scientific mode, which means structured ways of conducting the test, including structured methods of developing and asking the questions and relying on clearly defined decision rules in interpreting the outcomes.

The Comparison Question Test (CQT) that appears in several variations is by far the most common test in the field of polygraphy at least in specific-incident testing. Its main characteristic is that it includes direct questions about the relevant issues, known as the relevant questions (R) and other questions that are used for comparing the physiological reactions to the reactions that accompany the relevant questions. These questions are termed comparison questions (C).

[1] Main portions of the ideas and data presented in this paper were first introduced to the professional community by the author at the APA Annual Seminar held in Nashville TN, August 2009.

[2] Avital Ginton, Ph.D, is a Psychologist and a retired Commander (Colonel) of Israel Police, in which he served for 22 years (1977-1999). Dr. Ginton has been a full APA member since 2002, and the President of the Israel Polygraph Examiners Association during 1995 – 2001. He has authored several articles appearing in this and other publications.

In order to make a Deception Indicated (DI) call in the test, the measured physiological reactions to the relevant questions must be stronger than measured reactions to the comparison questions and vice versa, i.e. a call of Non-Deception Indicated (NDI) is contingent upon the occurrence of stronger reactions to the comparison questions compared to the relevant ones. Thus we may symbolize these very basic decision rules by:

R>C = Deceptive examinee
R<C = Truthful examinee

These basic rules which relate to the relative magnitude of the physiological reactions to the relevant vs. comparison questions are assumed to reflect the differential amount of concern that an examinee has towards the two categories of questions.

The core claim is that given a careful development and formulation of the questions, deceptive examinees are more concerned about the relevant questions relative to the comparison questions, while the opposite is true for truthful examinees, and this differential concern is manifested in the relative magnitude of the physiological reactions (Matte, 1996; Raskin and Honts, 2002). In order to measure the relative magnitude of the reactions, namely the difference found between the magnitude of the reactions that accompany the relevant questions with those that appear in the comparison questions, several sets of rules have been suggested, that in their more formalized version have been termed the Numerical Scoring Technique.

The Numerical Scoring Technique was introduced into the polygraph arena by Cleve Backster some fifty years ago (Backster, 1963). Basically it states that each time a relevant question is asked on the test, the physiological reactions that accompany the question and its answer, in each of the physiological measures are to be compared to the parallel reactions in an adjacent comparison question. These comparisons are conducted for each of the relevant questions in every repetition that takes place. In each comparison point a numerical score is given to represent the direction and the magnitude of the difference of the physiological reactions between the compared R and the C questions. Whenever the reaction shown in R question is stronger than the compared reaction in C, a negative number is given to that comparison point, and a positive number signifies a stronger reaction to the C question relative to the R question in that specific comparison point. Some numerical scoring techniques use a three-point scale of +1, 0 ,–1, but most of them, including Backster's original technique, use a seven-degree scale of +3,+2,+1,0,-1,-2,-3, which relates not only to the direction of the difference but also to its magnitude. Thus, a huge difference receives +/-3 while a just noticeable difference will get only +/-1. The numerical scoring techniques draw upon the fact that in all CQT versions the questions are asked more than once, so the individual scores can be added to reach a total sum. Usually the total must reach a certain cut score in order to make a call, otherwise the test is deemed inconclusive.

Thus we may say that, in fact, two separate factors are involved in the Numerical Scoring Techniques. The first one is the detected differences in reactions between R and C questions in specific comparison points and the second is the level of consistency found along the test with regard to these differences. The first factor is manifested in comparing the reactions to R and C questions in specific comparison points, and attaching numerical values to any discerned difference, while the second is expressed in checking for consistency of these differences, by repeating the act of comparison in other comparison points and adding the scores to a grand sum.

The common undeclared and perhaps unaware attitude towards the latter, i.e. the consistency factor, has been to relate to it as a secondary factor relative to the primary role given to the magnitude of difference in reactions between the R and C questions. This undeclared and perhaps unaware attitude reveals itself through the existing practice as can be shown in the following example:

If we take two 3-chart tests (a test with 3 repetitions of the series of questions), we may get the following numerical scorings:

Test A		**Test B**	
Ch1	- 8	Ch1	-2
Ch2	+1	Ch2	-1
Ch3	+1	Ch3	-1
------------------		------------------	
Total	-6	Total	-4

Most formal numerical scoring techniques will make a Deception Indicated (DI) call in the first case and deem the second case Inconclusive (INC), although in terms of consistency between charts, the second case rather than the first one is more indicative of deception since all three charts are pointing at the same direction, namely, it keeps showing that the reactions associated with the relevant questions are on average stronger than the comparison questions reactions, just as the theory would claim.

The theoretical rational of CQT states very clearly the expected direction of the differences between R and C questions; however, it says nothing about the size of these expected differences in the magnitude of reactions.

Why is it then that we are looking for a certain minimum in the magnitude of the difference before rendering a decision? It is because we want to make sure that the observed difference is reliable and not a mere reflection of random fluctuations of the psycho-physiological activity, or what we may call, ‘irrelevant noise.” Somehow we assume, probably justifiably, that small differences due to noise effect might occur frequently, as opposed to big differences, which are very rare to occur due to random fluctuations per-se. Hence we want to see a big enough difference before attaching any significance to it. While that might seem quite a reasonable approach, one should keep in mind that in order to avoid relying on an observed difference that occurred just by random fluctuations or other kinds of non-systematic noise, the consistency factor is much more important than the size of the difference, simply because that is what the consistency factor is all about.

It should be clarified that within the realm of testing theories, consistency is highly related to reliability. If a measurement device or procedure consistently assigns the same score to individuals or objects with equal values, the device or the procedure is considered reliable but if the scores assigned to the same individuals or objects vary in repeated measurements in the absence of any known and understandable reason, the device or the procedure is considered unreliable

Reliability is the extent to which a test or any measuring procedure yields the same results on repeated measures or trials, or the extent to which a repeated test yields consistent outcomes and scores. Thus, reliability can be defined by the consistency of a measurement procedure and its outcomes.

Consistency is not just a “nice to have” factor, rather it is a critical factor for decision making in CQT (and other polygraph techniques). We want to ensure that the outcomes we get on the test using our measuring procedure represent the true states of the examinees in their relative concerns about the relevant versus the comparison questions which define our decisions, and not a mere reflection of random fluctuations.

Acknowledging the importance of the consistency factor, the following is an attempt to introduce into the Numerical Scoring Technique a set of Decision Rules based mainly on consistency between charts. At this stage it has been applied only to single issue specific examinations.

The Between-Charts-Consistency Method of Analysis

In accordance with the basic rational of the CQT and the numerical scoring technique, the sum of scores in each chart can be negative, positive or zero. A negative number is pointing to Deception, a positive score points to a Truthful outcome and of course, zero doesn't have any inclination. In line with this, each chart is marked as a D chart, a T chart or a Z chart, and the span of possible combinations for 3-charts examination is as follows:

3D	- The sum of scores in each of the three charts is Negative.
3T	- The sum of scores in each of the three charts is Positive.
2Dz	- Two charts are scored Negative, and one is scored Zero.
2Tz	- Two charts are scored Positive, and one is scored Zero.
2Dt	- Two charts are scored Negative, and one is scored Positive.
2Td	- Two charts are scored Positive, and one is scored Negative.
1dZ	- One chart is scored Negative, the rest are scored Zero.
1tZ	- One chart is scored Positive, the rest are scored Zero.
1dtz	- One chart is scored Negative, one Positive and one Zero.
3Z	- The sum of scores in each of the three charts is Zero.

Remarks: a) D or d stands for Deception indicated trend; T or t stands for Truth telling indicated trend; Z or z stands for Zero or no trend at all.
b) Capital letter in the category name indicates majority and lower case, minority.
c) Chart order was not considered.

Given the above span of Between-Charts-Consistency categories, the following is a suggested set of decision rules based on these categories to be used in numerical scoring techniques. The adequacy of these decision rules has been tested on 1000 scorings of confirmed examinations, as will be shown later in this paper.

Decision Rules Based on Consistency-Between-Charts

Primary Rules

1. The calls are based on the direction in which the majority of the charts are pointing. Thus: A DI call is given in case of 3D, 2Dz or 2Dt. A NDI call is given in case of 3T, 2Tz, or 2Td

2. When the number of charts pointing at each direction are equal and no majority can be established, the call is INC. Thus an Inconclusive (INC) call is given in case of 1dtz. The same holds for 3Z.

Secondary Rules

3. In the case that two charts are scored Zero and only one is pointing at a certain direction, i.e. 1dZ or 1tZ, if the score is between +/-3 the call is INC and if it is at least +/-4 the call is either DI or NDI depending on the direction.

4. In the case of 2Dt or 2Td, if the grand sum score is pointing at a direction opposite to the direction pointed by the 2 majority charts, the call is INC.

Example: A 2Dt case: ch1 –2; ch2 +4; ch3 – 1; Grand Total = +1.

The positive (+) final score contradicts the direction suggested by the majority. The call is INC.

In order to test the compatibility of this set of decision rules, they were implemented on 100 confirmed single-issue field Zone examinations, 50 guilty and 50 innocents, from NCCA data base, that had been scored blindly by 10 experienced examiners to make 1000 scorings, for another research (Krapohl & Cushman, 2006). Details with regard to the scorers and the exact procedure they went through can be found in the original research.

Applying the Consistency-Based 4 Rules on the 1000 original scorings resulted in the following outcomes:

Decision Accuracy for Deceptive Examinees	- 83.40%
Decision Accuracy for Truthful Examinees	- 91.88%
Overall Accuracy w/o Inc	- 87.81%
Inconclusive Rate	- 9.80%

In Table 1 a comparison is presented between the accuracy and inconclusive rates achieved by applying the decision rules of three conventional approaches and the Between-Charts-Consistency method, on the same data. In the upper part of the table it can be seen that the overall accuracy rate achieved by implementing the Between-Chart-Consistency decision rules on the 1000 scorings fell short of only the accuracy achieved by using the Utah-like cut scores.[3] However, while the Utah-like grand total cut scores resulted in a 24.2% Inconclusive rate, the Consistency-Based Decision Rules resulted in only a 9.8% Inconclusive rate, a difference that is statistically significant at the level of 0.05 as the two 95% confidence intervals do not overlap. In that respect it also outperformed the Investigative Decision Rules that produced 19.8% Inconclusive rate. The results of the Consistency Based Decision Rules in overall accuracy and inconclusive rate are very similar to those which have been achieved by implementing the Evidentiary set of Decision Rules with some difference (though not statistically significant) in the balance between FP and FN rates. While in the Evidentiary set they seemed to be equal, in the Consistency set, the rate of FN was found to be two times the FP rate. That might suggest that the accuracy rate of identifying truth-tellers is higher when using the Consistency set at the expense of lower accuracy in detecting Deceptive examinees. Further refinements might change this imbalance, but whether to do it or not should be subjected to the philosophy (or policy) guiding the examiners or the organizations that conduct the examinations with respect to the relative cost of the two types of errors.

The Consistency Based Decision Rules can function as an alternative TDA model but the very fact that it uses some non-overlapping information makes it worthy to check whether a combined usage of it with, for instance, the Evidentiary Decision Rules improves the outcome beyond the level of each one of them as a stand-alone set. There are several options of how to combine them. One reasonable option is to implement a rule that in order to deem an exam inconclusive it must be found inconclusive by both sets of decision rules, or that there is a contradiction between the outcomes produced by the two different sets. It should be mentioned that contradictions can occur only when the second stage of the Senter Two-Stage TDA is affecting the call in the Evidentiary Decision Rules. In those cases the outcomes of some examinations turn out to be DI with Evidentiary set of rules while the Consistency Based Decision Rule indicates NDI. It is expected that contradictions will be rare events and in the present sample it took place only in 1.75 % of the outcomes.

[3] The difference only approaches statistical significance since there is a small overlap between the two 95% confidence intervals.

Table 1. Percent of decision accuracy, inconclusive rates and confidence intervals reached by different sets of decision rules in 1000 scorings*. (in brackets - 95% confidence intervals for proportions)

Dec Rules		DECEPTIVE			TRUTHFUL			OVERALL	
		Inc Rate	Correct w/o Inc	Sensitivity	Inc Rate	Correct w/o Inc	Specificity	Inc Rate	Correct w/o Inc
Investigative ** Decision Rules		13.2 (+/-3.0 10.2 to 16.2)	94.9 (+/-2.1 92.8 to 97.0)	0.824 (+/-0.033 0.791 to 0.857)	26.4 (+/-3.9 22.5 to 30.3)	75.5 (+/-4.4 71.1 to 79.9)	0.556 (+/-0.044 0.512 to 0.600)	19.8 (+/- 2.5 17.3 to 22.3)	86.1 (+/- 2.4 83.7 to 88.5)
Evidentiary ** Decision Rules		8.6 (+/- 2.5 6.1 to 11.1)	86.7 (+/- 3.1 83.6 to 89.8)	0.792 (+/- 0.036 0.756 to 0.828)	6.0 (+/- 2.1 3.9 to 8.1)	87.7 (+/- 3.0) 84.7 to 90.7)	0.824 (+/-0.033) 0.791 to 0.857)	7.3 (+/- 1.6) 5.7 to 8.9)	87.2 (+/- 2.2 85.0 to 89.4)
Utah Like*** Grand Total Cut-Scores Inc= +/- 5		28.0 (+/-3.9 24.1 to 31.9)	87.2 (+/- 3.5 83.7 to 90.7)	0.628 (+/-0.042 0.586 to 0.670)	20.4 (+/-3.5 16.9 to 23.9)	95.4 (+/- 2.0 93.4 to 97.0)	0.760 (+/-0.037 0.723 to 0.797)	24.2 (+/- 2.7 21.5 to 26.9)	91.6 (+/- 2.0 89.6 to 93.6)
Between-Chart-Consistency Decision Rules		13.0 (+/-3.0 10.0 to 16.0)	83.4 (+/- 3.5 79.9 to 86.9)	0.726 (+/-0.039 0.687 to 0.765)	6.6 (+/-2.2 4.4 to 8.8)	91.9 (+/-2.5 89.4 to 94.4)	0.858 (+/-0.031 0.827 to 0.889)	9.8 (+/-1.8 8.0 to 11.6)	87.8 (+/- 2.1 85.7 to 89.9)
Change in unit of analysis*									
Consistency Based Decision Rules applied on distinct examinations (100) as the units for analysis rather than distinct scorings (1000)	100 examinations each scored by 10 scorers Index #1*	12.0 (+/-9.0 3.0 to 21.0)	86.4 (+/-10.1 76.3 to 96.5)	0.760 (+/-0.118 0.642 to 0.878)	4.0 (+/-5.5 0.00 to 9.5)	95.8 (+/- 5.7 90.1 to 100.00)	0.920 (+/-0.053 0.867 to 0.973)	8.0 (+/- 5.3 2.7 to 13.3)	91.3 (+/- 5.8 85.5 to 97.1)
	100 examinations each scored by 10 scorers Index #2*	12.0 (+/-9.0 3.0 to 21.0)	88.7 (+/-9.4 79.3 to 98.1)	0.780 (+/-0.115 0.665 to 0.895)	2.0 (+/-3.9 0.00 to 5.9)	95.9 (+/-5.5 90.4 to 100.00)	0.940 (+/-0.066 0.874 to 1.00)	7.0 (+/- 5.00 2 to 12)	92.5 (+/-5.4 87.1 to 97.9)

* See text for further clarifications.
** Data from Krapohl, & Cushman (2006).
***Note that it is not the complete Utah method of analysis but only the use of its decision cut scores.

Interestingly enough, in two-thirds of them the results reached by the Consistency Based Decision Rules were correct and only one-third of the Evidentiary TDA were so. In cases that are judged Inconclusive by only one set of rules, the calls should follow the other set's verdict. When applying this option to the present sample the inconclusive rate was reduced to 5.5%, a decrease of 25% (chi^2 =6.811; df=1; p<0.01)[4], and the accuracy of decisions was 87.3% (increase of 0.1%) relative to the Evidentiary TDA as a stand-alone method. This is but a small improvement in the Inconclusive rate that demonstrates in principle the potential contribution of the consistency factor to the outcomes. Acting similarly on the Investigative TDA outcomes demonstrates an even higher impact resulting in a decrease of 48% in the Inconclusive rate (chi^2 =46.29; df=1; p<0.0001) and the accuracy of decisions was improved by roughly 2 % (n.s.).

It is interesting to look at the data as it is distributed between the various categories of the Between-Charts-Consistency.

Table 2. Distribution of 1000 scorings in the Between-Charts-Consistency categories

Category	Frequency
3T	274
3D	204
2Tz	92
2Dz	47
2Td	150 (134)
2Dt	157 (143)
1tZ	6 (2)
1dZ	9 (6)
1dtz	60
3Z	1

In brackets - w/o Inc.

Assuming the basic premise of the CQT that R > C indicates deception and R < C indicates truthfulness with regard to the relevant questions is correct, the rational of the Between-Charts Consistency Decision Rules predicts different accuracy rates amongst the various categories, i.e. in categories that indicate high between-chart consistency, the accuracy rate should be higher than in categories with low between-chart consistency. The data with this regard is presented in Table 3.

[4] McNemar test for the significance of change. When the Yates correction for continuity was added the result was Chi^2 = 6.11; p<0.02, df=1

Not surprisingly, very high rates of accuracy are found in the categories that indicate a high consistency between charts (3T,3D,2Tz,2Dz) – a weighted average of 95% and only 75% in categories that indicate low consistency between charts (2Td,2Dt,1tZ,1dZ). The difference in accuracy of decisions between these two levels of consistency categories is highly significant (Chi^2 = 79.01, df=1, p< 0.0001). That might have some implications for the kind of decisions made by the examiner. In the Israeli police it is customary to use two levels of confidence in the results that are handed to the investigation unit by the examiners, a definitive and a reserved result.[5] In line with this approach and vis-à-vis the findings presented in this paper, it is recommended to go for a definitive result when high between-chart-consistency has been identified, and opt for a reserved result in the case of low between-chart consistency.

Table 3. Accuracy of decisions per category, based on Between-Chart Consistency, in three-chart examinations[a]

Consistency Category	Accuracy Rate w/o Inc	Type of Error	Portion of the Scored Sample	Number of 3-Chart Exams Scorings	Inc Rate
3T	97.45%	FN	27.4%	274	0
3D	96.08%	FP	20.4%	204	0
2Tz	84.78%	FN	9.2%	92	0
2Dz	95.74%	FP	4.7%	47	0
2Td	65.67%	FN	15%	150(134)[b]	10.6%
2Dt	81.81%	FP	15.7%	157(143)[b]	8.9%
1tZ	100%	FN	0.6%	6(2) [b]	66.7%
1dZ	83.33%	FP	0.9%	9(6) [b]	33.3%
1dtz	------------	--------	6.0%	60	100%
3Z	------------	--------	0.1%	1	100%

[a] 100 confirmed single-issue field Zone examinations, 50 guilty and 50 innocents, scored by 10 experienced examiners to make 1000 scorings. From DACA data base.
[b] In parentheses, w/o Inc.

[5] In the Israel police approach which is more flexible than the federal approach and has been termed internally the Israel Police Modular Technique (IPMT), (Ginton & Ber, 1992), the practice is that the examiners are entitled to make a definitive call or a reserved one which is relatively free from the rule of rigid cut scores or any rigid alpha levels. Thus, if the final outcome doesn't reach the cut score it is not automatically deemed Inconclusive. The examiner can choose to opt for a reserved call that reads in free translation, "The outcome was not conclusive but inclination toward DI/NDI was observed." Likewise, a final outcome that exceeds the cut score doesn't automatically results in a conclusive verdict and the examiner is entitled to opt for a reserved call such as "The outcome points towards a DI/NDI verdict but it should be treated cautiously in a reserved manner due to.......”

The above analysis was based on 1000 scores, however in fact there were only 100 examinations (conducted on 100 examinees) each scored by 10 scorers. Thus the situation bears a factor of consistency between scorers (inter-judges reliability). A perspective that put the examinations (100) as the main units for analysis rather than the individual scorings (1000) seems to be important for further testing the viability of the decision rules based on Consistency-Between-Charts. That can be done by combining the 10 Between-Charts-Consistency outcomes in each examination into one figure that indicates across the variability found among the scorers whether the results of a certain examination is inclined towards Deception or Non-Deception Indication. This analysis incorporates on top of the Between-Charts-Consistency also the degree of consistency achieved between scorers. To accomplish it, two alternative indices were developed as follows:

Index #1 - Based on the Between-Charts-Consistency Decision Rules, each examination got 10 decisions derived from the 10 scorings (one per scorer). Each one of them could be NDI, DI, or Inconclusive. Index #1 consisted of the percentage of scorers with NDI Calls, less the percentage of scorers with DI Calls (%NDI – %DI) per each examination. The decision for each examination, beyond the variability between the scorers, was reached by implementing the following decision rules on the Index Outcomes.

Index Outcomes between +/- 20% = INC
Index Outcomes of 30% or higher = NDI
Index Outcomes of -30% or lower = DI

Two hypothetical examples might help understanding the Index. Suppose a certain examination, based on the Between-Charts-Consistency Decision Rules, was deemed NDI by six scorers (60%), INC by 3 scorers and DI by 1 scorer (10%), the examination Index Outcome is 60%-10% = 50% and the overall outcome of this examination is NDI. In another case the examination was deemed NDI by six scorers (60%), INC by one scorer and DI by four scorers (40%), the examination Index outcome is 60%-40% = 20% and the overall outcome of this examination is INC.

The alternative index, Index #2 - Each category of Between-Charts-Consistency received a numerical score relative to the degree of consistency and its sign as follows:

3T	=	+3
2Tz	=	+2
2Td	=	+1
1tZ	=	+0.5
1tdz	=	0
3Z	=	0
1dtz	=	0
1dZ	=	-0.5
2Dt	=	-1
2Dz	=	-2
3D	=	-3

The numerical consistency scores given by the 10 scorers in each examination were added to reach a total score between 30 to -30. The decision for each examination, beyond the variability between the scorers, was reached by applying the following decision rules on the total outcomes:

Decision rules:

Index Outcomes between +/- 2 = INC
Index Outcomes of 2.5 or higher = NDI
Index Outcomes of -2.5 or lower = DI

The inconclusive zones for both indexes were set to optimize the outcomes, and of course should be checked for adequacy with more cases. To further clarify the Indices procedures, a sample of combined outcomes of consistency scores of 10 scorers and the two alternative Indices are presented in Table 4.

Table 4. Examples of outcomes of consistency scores of 10 scorers and overall combined outcomes with two alternative indices

		SCORERS						Values of the two indexes & Final Calls Per Exam	
Exam#	Gnd	#1	#2	… …	#9	#10	Frequencies of Consistency Scores	Sum Inx 1	Sum Inx 2
01	T	2Td	3T	…	3T	2Tz	10NDI	100 NDI	24 NDI
02	T	3T	3T	…	2Dt	2Dt	4NDI,5DI,1INC	-10 INC	1.5 INC
03	T	1tdz	3T	…	2Tz	1tdz	5NDI,5INC	50 NDI	11 NDI
04	T	2Td	2Td	…	2Td	2Dt	6NDI,3DI,1INC	30 NDI	3 NDI
05	T	2Tz	2Tz	…	2Td	2Td	10NDI	100 NDI	13 NDI
06	T	3D	1dtz	…	3D	1dtz	1NDI.5DI,4INC	-40 DI	-9.5 DI
07	T	3T	3T	…	3T	3T	10NDI	100 NDI	30 NDI
51	D	2Dz	3D	…	2Dz	2Dt	1NDI,8DI,1INC	-70 DI	-24 DI
52	D	2Dt	2Td	…	2Td	2Td	8NDI,2DI	60 NDI	5 NDI
53	D	2Tz	2Dt	…	2Td	1tdz	4NDI,4DI,2INC	0 INC	1 INC
54	D	3D	2Dt	…	3D	3D	10DI	-100 DI	-24 DI
55	D	3D	3D	…	3D	3D	10DI	-100 DI	-30 DI
56	D	2Td	1tdz	…	1tdz	1tdz	2NDI,3DI,5INC	-10 INC	-3 DI

Hit	Miss	Inc

The following are the outcomes based on these two indices of 100 confirmed examinations (50 deceptive and 50 truth-teller examinees), percentage of accuracy and INC rate:

Computed with Index #1

50 Truthful Examinees:	46T, 2D, 2 INC
Accuracy (w/o Inc):	95.83%
Specificity:	92.00%
50 Deceptive Examinees:	38D, 6T, 6 INC
Accuracy (w/o Inc):	86.36%
Sensitivity:	76.00%

Unweighted Mean accuracy (w/o INC): 91.095%
Overall Accuracy (w/o INC): 91.30%
Inconclusive Rate: 8.00% INC = +/- 20%

Outcomes based on between charts and beyond scorers consistency of 100 confirmed examinations (50 deceptive and 50 truth-teller examinees), percentage of accuracy and INC rate. computed with Index # 2

50 Truthful Examinees:	47T, 2D, 1 INC
Accuracy (w/o INC):	95.92%
Specificity:	94.00%

50 Deceptive Examinees: 39D, 5T, 6 INC
Accuracy (w/o INC): 88.64%
Sensitivity: 78.00%

Unweighted Mean accuracy (w/o INC): 92.28%
Overall Accuracy (w/o INC): 92.47%
Inconclusive Rate: 7.00% INC = +/- 2

In Table 1 the two bottom rows in the lower part of the table indicate the accuracy reached by using these two indices. Comparing the figures reached by the various decision rules presented in the table shows that these indices, which are based on consistency between charts beyond the variability of the scorers, lead to the best results when considering both the very low Inconclusive percentage and the very high accuracy rates.

The decision rules based on Consistency-Between-Charts do not change the scoring rules that determine the numerical outcomes, rather it deals with changing the decision rules by which the numerical outcomes are evaluated prior to making a call.

However, another possible way to increase the weight of the consistency factor might be to introduce into the scoring system a consistency index that interacts with the existing numerical scoring rules to change the numerical outcomes while keeping the decision rules whether federal Investigative, Evidentiary, Utah or others, unchanged. An example of this approach was presented by the author in the APA annual seminar (Ginton, 2009)[6]. In that example when the consistency index was implemented on the 1000 scores in the afore-mentioned sample, the Inconclusive rate was improved (lowered) by 26%, while using the Utah like Grand Total cutoff points, at the expense of reducing accuracy of decisions by only 1%.[7]

It should be mentioned that due to the lack of any research on this issue, the consistency index was not supported by any previous published data, rather it was developed through many years of experience and trial and error explorations. The exact procedure or the numerical factor (0.5) used in that index is of a little importance, what seems to be important is the notion of introducing into the numerical scoring techniques a way to give the consistency factor its appropriate weight.

More about the meaning of Reliability/ Consistency and their implications to polygraph testing

Another way to grasp the meaning of Reliability/Consistency has to do with the definitions of variability and test variance. A variable is any factor, trait, or condition that can exist in differing amounts or types, something that is subject to variations, and variability is the quality of being subject to variation. Test variance represents variability on how people perform on a test, a measure of the total amount of variability found on the test. When there is little variability, the variance is small, while a big test variance indicates a lot of variability found on the test.

[6] The suggested consistency index in that research was based on the following premises and rules:

A. When there are no "Zero" charts, count the number of charts that indicate the same direction (Plus or Minus). Each time it points to that direction add 0.5 to the index and multiply the scores that go in this direction by its final index value. Thus, if all three charts are in the same direction, multiply the scores by 1.5. If only two are in the same direction, multiply the scores which go in this direction by 1.0 (no change), and the score of the only chart that indicates an opposite direction, by 0.5. Thus its opposite effect is reduced by half.

B. In case there are more than three charts, the same rule of adding 0.5 to the index per chart is applied, thus, the multiplication factor can reach N/2 when N is the number of charts per exam.

C. When "Zero" chart exist, if a majority of the other charts point to the same direction, the "Zero" chart is considered supportive by not contradicting the majority trend and adds also 0.5 to the index. In other cases the original scores are not changed.

[7] Results of implementing this consistency Index on the 1000 scores while using Utah like Grand Total cutoff points for the inconclusive zone changed the Inconclusive rate from 24.2% to 17.9% and overall correct decisions w/o inconclusive from 91.55% to 90.38%. Sensitivity and specificity were increased by almost 5%.

This variability may represent true variations between the people in the quality, amount or degree that they hold in that particular variable. For instance, some people are smarter than others, but it may also be the outcome of irrelevant noise or fluctuations due to measuring inadequacies or effects of irrelevant factors. For instance, some people were more tired than the others while taking the test or in the last 10 minutes of the test a sub group of the students have been exposed to irritating sun beams entered through the window.

In line with this notion we can find the "True Score Theory" (Lord, Novick & Birnbaum, 1968; Trochim, 2000) that in its simplified version, when adapted to our field, states that - the measured reactions to any question and the measured difference between the strength of reactions to relevant vs. comparison questions, are a combination of "True score" plus "Error". That is to say that in principle the output itself always contains irrelevant aspects and the measurement device or the act of measuring might add another irrelevant value, namely Measurement Error. Thus the polygraph test variance includes true variability in the values held by the examinees plus variance that reflects the effect of irrelevant factors.[8] A person's test score might deviate from the true score because he/she was sick, anxious, in a noisy room, inattentive, got stomach cramps, distracted by examiner's inappropriate behavior, etc.

A main portion of these irrelevant factors is the existence of random fluctuations, for instance in the momentary level of the examinee's attention capacity or focus and in the pattern and the strength of the psycho-physiological reactions. These fluctuations might mask the true values and interfere with making the right decision. The need to nullify or at least to weaken the effect of these fluctuations is crucial for any test and a polygraph test is no exception. Repeating the questions and looking for consistency over the repetitions is the main way to achieve that. The logic behind this tactic is that random fluctuations tend to nullify themselves in repeated measures and enable the true score to surface. The degree to which a certain technique or test is immune to the effects of such fluctuations on the outcomes is defined by the Reliability index that takes values between 0.0 to 1.0, indicating the extent to which an assessment tool or procedure is consistent, i.e. free from random error in measurement.

As already mentioned, the CQT is based on the assumption that deceptive examinees are more concerned about the relevant questions relative to the comparison questions, while the opposite is true for truthful examinees, and that these differential concerns are manifested in the relative magnitude of the physiological reactions. The differences found between the magnitude of the reactions that accompany the relevant questions and those that appear in the comparison questions indicate the veracity of the examinee with regard to the relevant questions. Looking for basic consistency in that matter means repeating the measuring of the difference in reactions between relevant and comparison questions and finding out whether the detected difference keeps pointing at the same direction.

Having stressed the crucial role of repetitions one may wonder about the number of repetitions needed for gaining the confidence that the outcome is reliable. Most techniques would ask for three repeated tests or three charts. Is that enough? Are three charts enough to satisfy the need for consistency when looking from the Consistency Factor perspective?

[8] It is important to understand that the variable here is not deception vs. truth-telling rather it is the difference between the reactions to the relevant and the comparison questions that supposed to reflect the difference in concerns between the questions. This difference can take various values which are the combination of true score (i.e. reflecting the amount of concern) plus error. There are two kinds of errors, random and systematic. Striving for consistency deals with the effort to reduce random error but can not remedy the effect of possible systematic error.

To make it simple for the moment, there are two options; the examinee is either Deceptive or Truthful. The possible influence of Chance fluctuations alone on the outcome is equal for both options, a situation that percentage wise we may call "Fifty-Fifty" (but, see later).

Let us take a look at a pure "Fifty-Fifty" situation – flipping a coin.

Chances of a proper coin to keep falling on heads in a series of tosses

1. 0.5
2. $0.5^2 = 0.25$
3. $0.5^3 = 0.125$
4. $0.5^4 = 0.0625$
5. $0.5^5 = 0.0312$
6. $0.5^6 = 0.0156$
7. $0.5^7 = 0.0078$
8. $0.5^8 = 0.0039$

It is clear that repeating questions or charts is not the same as repeating coin tossups. For one thing, in polygraph testing the repetitions are in a way dependent, meaning that the experience the examinee accumulates from each repetition might affect his attitude towards the next repetitions and probably also the sort and strength of the reactions, while in the tossing series, the outcome of each toss is totally independent of previous tosses. Because we do not know for sure how each repetition in polygraph testing affect the next ones we may assume, for the sake of estimating the chance effect, independency between the repetitions of questions or charts and get the same probability table of chance outcomes that is presented above also for the direction of the differences between strength of reactions to relevant vs. comparison questions i.e. positive or negative scores.

Probability of charts outcomes to point in the same direction (positive or negative) on different numbers of successive charts by chance only (ignoring zeros)

1. 0.5
2. $0.5^2 = 0.25$
3. $0.5^3 = 0.125$
4. $0.5^4 = 0.0625$
5. $0.5^5 = 0.0312$
6. $0.5^6 = 0.0156$
7. $0.5^7 = 0.0078$
8. $0.5^8 = 0.0039$

It is only a crude estimate but we can see that if we get for instance, on each of three successive charts, R>C, there is still a probability of 12.5% that it is a chance effect per-se, let alone cases in which only two of the three charts pointing at the same direction. In order to evaluate the meaning of this probability on polygraph testing, I would like to turn for a while to what is known as inferential statistics.

When trying to prove statistically the existence of certain phenomenon or effect, we use an inferential statistics approach in which we formulate two opposing statistical hypotheses: the Null Hypothesis (H_0) and the

Alternative Hypothesis (H_1). H_0 postulates that the effect doesn't exist in reality and the measured value reflects only chance fluctuation. Contrary to that, H_1 states that the effect shown by the measured value is a real one.

The two hypotheses are tested by asking: what is the probability of getting the measured value by pure chance? If the outcome of the chance figure is very low, one may reject H_0 and conclude that it is conceivable that the found value or the result is not a chance case but rather reflects a true score and a true effect.

Getting back to the Consistency-Between-Charts issue, we can use inferential statistics to formulate H_0 that states that the effect shown in a test is but a chance effect and an alternative hypothesis H_1 that states that the effect found is a real one. Given that from the Consistency-Between-Charts perspective, in a three-chart examination the probability of having all three of them pointing at the same direction by mere chance is 0.125 (12.5%), we should ask ourselves is this probability low enough to permit a rejection of H_0 and render a reliable call?

Unfortunately, the convention in the behavioral sciences is that "alpha", the probability associated with getting certain values by chance, should be 5% or lower in order to reject H_0 and turn to H_1, meaning that the 12.5% associated with chance probability of getting three charts consistency is not scientifically enough to allow a reliable call. From this perspective it means that even if it is found for instance in three consecutive charts that R>C, the maximum that one can count on to believe that the shown direction reflects a true quality of the examinee rather than random error does not exceed the level of 87.5% of confidence.

Now, due to the diagnostic capability of the test, the actual situation is not a "Fifty-Fifty" one. The assumed accuracy of a single chart is in the neighborhood of 80% (estimated from the proportions of charts pointing at the correct vs. incorrect directions in the current sample, see Table 5) which means that ignoring Zero charts, there is a chance of 20% to get a chart score that points at the wrong direction. Based on this estimate, the probability of getting three out of three charts pointing at a wrong direction by chance is $0.2^3=0.008$, which is well below the alpha convention of 5% and therefore in that case we can rely on the outcome as pointing at the right direction, i.e. being correct. However, in the case that one of the three charts is pointing to the other side and only two out of three are pointing at the same direction, the probability that the outcome of the test is associated with chance increased dramatically ($p>0.1$) not allowing rejection of H_0.

This shortage is assumed to be resolved by the common numerical scoring techniques that minimize the weight of the consistency factor, relying only on the aggregated scores. But as has been suggested throughout this paper the present author believes that giving higher weight to the consistency factor will result in improving the outcomes.

Does this mean that we should run more than three charts as a default, perhaps four or five? The Utah approach suggests scoring the test after three charts and in case the outcome is inconclusive to go for another two charts (Raskin & Hont, 2002). This suggestion has been lately adopted by others (Federal Psychophysiological Detection of Deception Examiner Handbook, 2011). The difference between Utah's suggestion and the suggestion derived from the above argumentation is that in the current suggestion, running four or five charts should be the default procedure irrespective of the outcomes after three charts. It means that sometimes conclusive outcomes reached after three charts might be changed to inconclusive after the forth or the fifth chart should these extra charts point to the opposite direction. This is pity of course in the case that the outcome after the third chart was a hit but it might also prevent a mistake when the outcome after three charts was an error. The trade-off of these two possibilities should be explored in research. As for now, from the perspective that stresses the importance of the Consistency Factor, and based on conventions used by the scientific community for dealing with chance effect, there is an apparent need to collect more than three charts. In the Israel police approach, which is less rigid than the federal approach and has been termed Israel

Table 5. Frequencies distribution of individual charts outcomes scored by 10 examiners in 100 three-chart examinations [a]

Categories of Between-Charts-Consistency	Number of Three-Chart Examination Scorings	Number of Individual Charts with Zero Outcome	Number of Individual Charts with Non-Zero Outcomes	Number of Individual Charts Pointing at the Correct Direction	Number of Individual Charts Pointing at Erroneous Direction
3T	274	0	822	801	21
3D	204	0	612	588	24
2Tz	92	92	184	156	28
2Dz	47	47	94	90	4
1dZ	9	18	9	6	3
1tZ	6	12	6	3	3
2Td	150	0	450	243	207
2Dt	157	0	471	280	191
1tdz	60	60	120	60	60
3Z	1	3	0	0	0
SUM	1000	232	2768	2227	541

[a] 100 confirmed single-issue field Zone examinations, 50 guilty and 50 innocents, scored by 10 experienced examiners to make 1000 scorings. From NCCA database.

Police Modular Technique (IPMT), (Ginton & Ber, 1992) it is customary to run three or four charts in which the last one is a double chart meaning that there are four to five repetitions of all questions.

To make a final point, consistency is one face of reliability, while accuracy of decisions is one face of validity. What can we learn from the Consistency factor about the Validity of the test? As said above, reliability can be defined as consistency in the type of result a test yields or the extent to which an assessment tool is consistent or free from random error in measurement. However, usually the reliability of a test is computed by measuring lots of examinees, and the reliability index as well as validity index relate to an overall quality of a certain test or technique and not to a single examination, while the Consistency-Between–Charts relates to any individual examination. Nevertheless, since in essence reliability has to do a lot with consistency, one may stretch the meaning of reliability to say that from the Consistency Factor perspective, because a perfect consistency in the direction of the difference found between R and C questions in three repeated charts still contains a 12.5% chance for random effect, or it is only 87.5% free from random error, even if we witness a full reliability between the charts' directions, the maximum that one can count on that to believe that the shown direction reflects a true solid direction rather than random error does not exceed the level of 87.5% of confidence.

What are the implications of the above for estimating the accuracy of the test? Accuracy is a common term replacing the more formal term - criterion validity - which is used in testing theory. In theory of testing it is well documented that the maximum validity that a test can reach is limited to the square root of its reliability. If we take it one step further it means that the degree to which one can count on the test to be free from random error, and sets a limit to the degree he/she can count on the found accuracy of the test to be a true reflection of reality in the relevant variable. In line with this logic, although 87.5% does not indicate formal reliability of the test, I take the liberty to say that from the Consistency Factor perspective the degree we can count on the accuracy figures reached from a three repeated test (three charts) might be limited to the square root of 0.875 which is 0.935.[9] It suggests that even in the case that the notion of differential reactivity to R vs. C questions, between the deceptive and the truthful examines, is absolutely correct, the maximum accuracy expected with three charts examinations is limited to 0.935. It should be pointed out that this indicates the limit for the overall validity in three-chart cases and not the limit in any particular sample or particular category of Between-Charts-Consistency and indeed the accuracy rates found with the current sample, in certain categories exceed this limit (see Table 3).

Two Words of Caution

The present demonstration of the viability of the Between-Charts-Consistency Decision Rules was done with verified examinations that had been conducted under the federal doctrine of how an exam should be run and scored. To enable generalization of this idea it should be replicated with verified examinations that have been conducted under different doctrines. For instance, while the common attitude under the federal approach is that no significant talking or stimulation should take place between charts, the Utah approach recommends it to a very high degree (Honts, 1999). This difference might be crucial to the issue of Between-Charts-Consistency.

Just before the end, there is one small reminder and reservation to be made. The unit of repetition in this Consistency Measure Application was the chart. Changing it to a question unit, while still be valid in principle, might result in different outcomes.

[9] In the same vein the degree one can count on the accuracy figure reached from a four-chart test is limited to the square root of 0.9375 which is 0.9682 and from five-chart test to the square root of 0.9688 which is 0.9842.

References

Backster, C., (1963). The Backster chart reliability testing method. *Law and Order*, 1, 63-64.

Federal Psychophysiological Detection of Deception Examiner Handbook (2011). *Polygraph*, 40, (1).

Ginton, A., & Ber, Y. (1992). *The polygraph doctrine – Understanding polygraphy - Chapters in theory and practice*. Unpublished Internal Manuscript (in Hebrew). Scientific Interrogation Lab, Israel Police.

Honts, C. R. (1999). The discussion of questions between test repetitions (charts) is associated with increased test accuracy. *Polygraph*, 28(2), 117-123.

Krapohl, D.J., & Cushman, B. (2006). Comparison of evidentiary and investigative decision rules: A replication. *Polygraph*, 35(1), 55-63.

Lord, F. M., Novick, M.R. & Birnbaum, A. (1968). *Statistical theories of mental Test scores*. Oxford, England; Addison-Wesely.

Matte, J. A. (1996). *Forensic Psychophysiology using the polygraph*. JAM Publications. Williamsville N.Y.

Raskin, D.C., & Honts, C. R. (2002). The comparison question test. In M. Kleiner, (Ed.) *Handbook of Polygraph Testing* (pp. 1-48). New York: Academic Press.

Trochim, William M.K. (2000). *The Research Methods Knowledge Base*, 2nd Edition. Atomic Dog publishing, Cincinnati, OH.