# Practical Policy Considerations on Internet Test Security and Score Interpretation - Cheating Candidates or Candidates that Cheat?

Rainer Kurz (rainer.kurz@savilleconsulting.com),
Rab MacIver, Peter Saville & Heidi Oxley (Saville Consulting);
Paul Roscorla (The Home Office)

BPS Occupational Psychology Conference

9th January 2008

# Presentation Outline

savilleconsulting

- Unsupervised internet testing
- Safeguard of supervised version
- Verification or Decision-making?
- Test-Retest Data modelling
- Swift Aptitude Assessment Overview
- Case study
- Q & A

# *Concordance Simulations*



4300 simulated cases to look at score differences expected without cheating. Two normally distributed variables were created which correlated .40, .70 and, .84 with each other. The random error on each of these variables was also normally distributed.

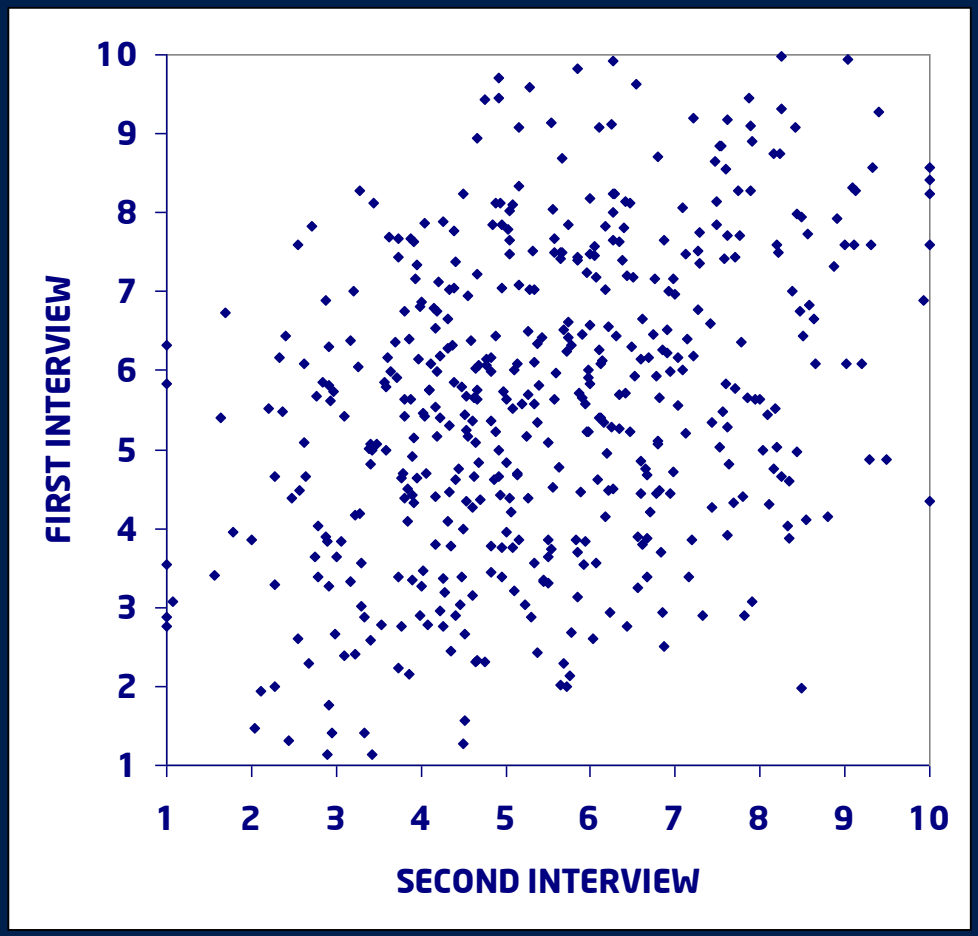
What are the Sten Score differences across two occasions:

- .40 correlation e.g. unstructured interview
- .70 correlation e.g. typical aptitude test
- .84 correlation e.g. longer aptitude test
- How can we distinguish between statistical random error and cheating?

# *Concordance Simulation with .4 Correlation*

savilleconsulting

## Concordance Between Scores at First and Second Interview (1-10 sten scale)



# *Concordance Simulation with .4 Correlation*

savilleconsulting

18% candidates have *no score change* between Interview 1 and Interview 2

16% of scores *drop* by 1 sten

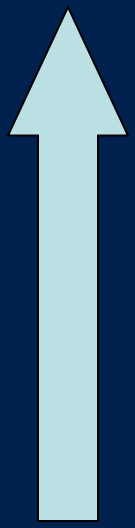
11% of scores *drop* by 2 stens

8% of scores *drop* by 3 stens

3% of scores *drop* by 4 stens

2% of scores *drop* by 5 stens

1% of scores *drop* by 6+ stens



16% of scores *increase* by 1 sten

11% of scores *increase* 2 stens

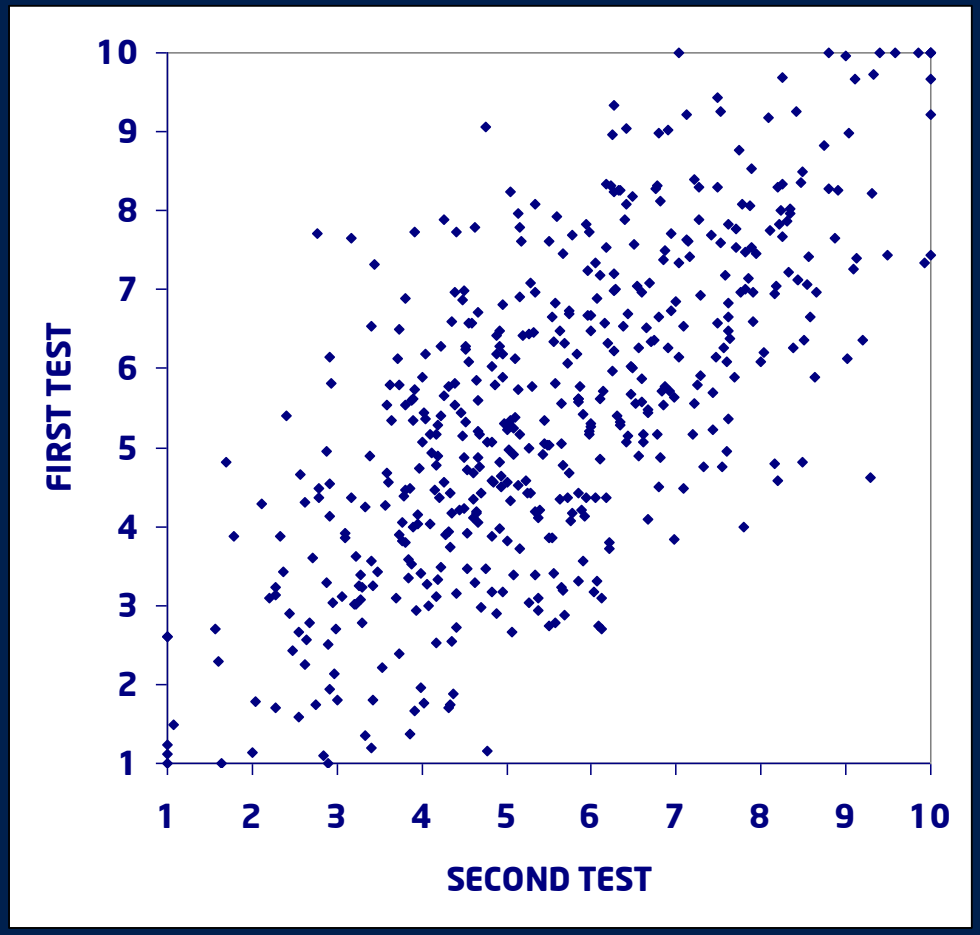
8 % of scores *increase* by 3 stens

3% of scores *increase* by 4 stens

2% of scores *increase* by 5 stens

1% of scores *increase* by 6+ stens

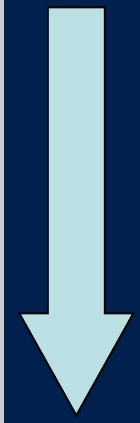
## Concordance Between Scores at First and Second Test (1-10 sten scale)



# *Concordance Simulation with .7 Correlation*

savilleconsulting

26% candidates have *no score change* between First and Second Test

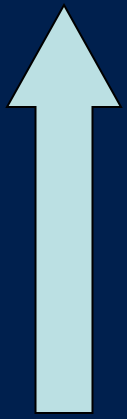


21% scores *drop* by 1 sten

10% scores *drop* by 2 stens

4% scores *drop* by 3 stens

2% scores *drop* by 4+ stens



21% scores *increase* by 1 sten

10% scores *increase* 2 stens

4% scores *increase* by 3 stens

2% scores *increase* by 4+ stens

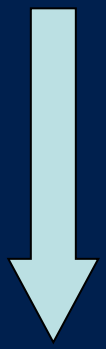
# *Concordance Simulation with .84 Correlation* savilleconsulting

## Concordance Between Scores at First and Second Test (1-10 sten scale)

# *Concordance Simulation with .84 Correlation*

savilleconsulting

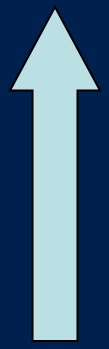
34% candidates have *no score change* between First and Second Test



22% scores *drop* by 1 sten

9% scores *drop* by 2 stens

2% scores *drop* by 3+ stens



22% scores *increase* by 1 sten

9% scores *increase* 2 stens

2% scores *increase* by 3+ stens

# *Concordance Simulations*

savilleconsulting

If Sten Score differences > 2 is 'verification' criterion:

- .40 correlation e.g. unstructured interview: 28% False Positives
- .70 correlation e.g. typical aptitude test: 12% False Positives
- .84 correlation e.g. longer aptitude test: 4% False Positives
- We cannot distinguish between statistical random error and cheating!

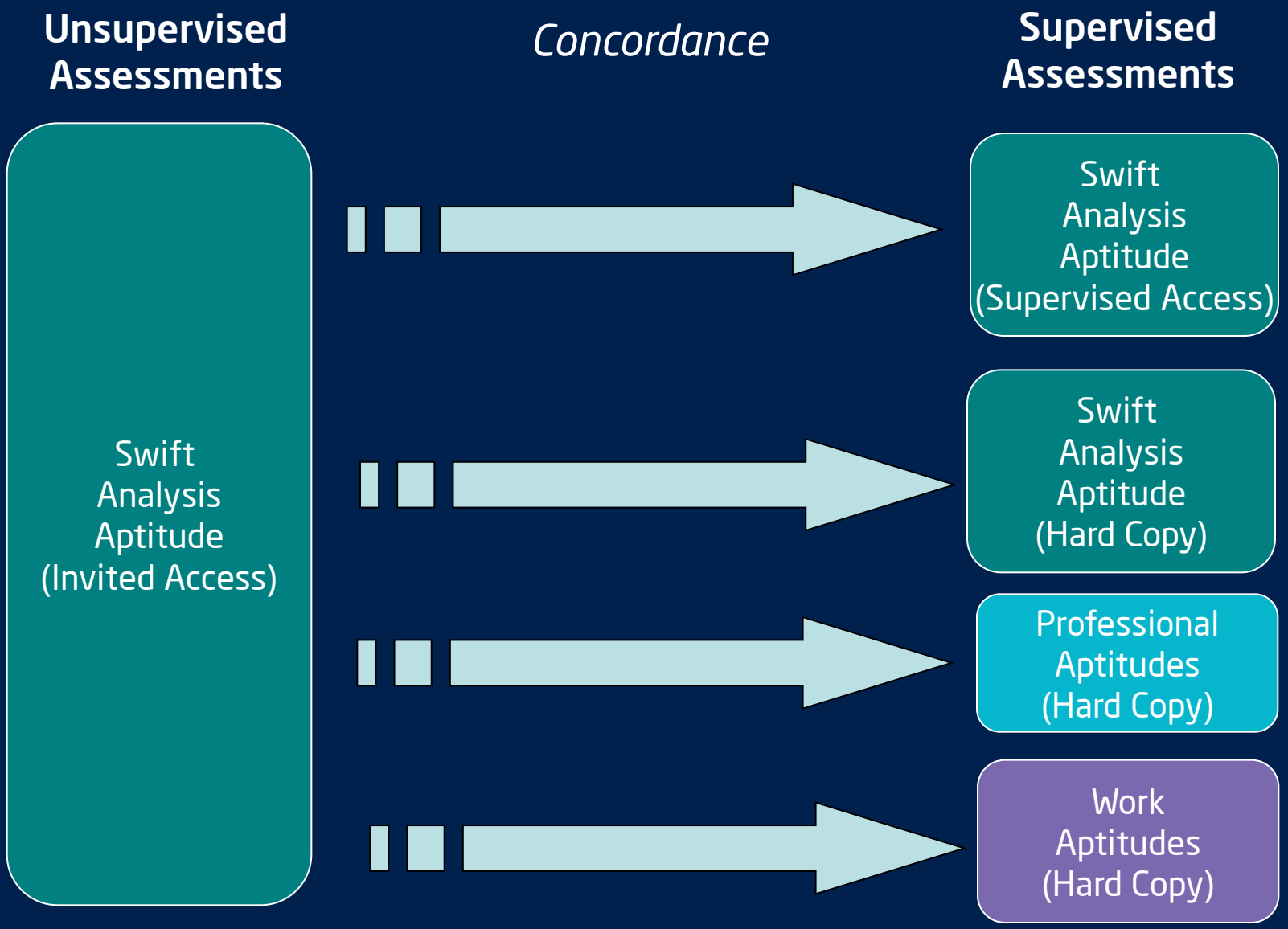
# Aptitude Assessment Portfolio

savilleconsulting

| Separate Aptitude Tests | Target Groups | Combination Assessments |
|---|---|---|
| Professional | Managers, directors and professionals. | Swift Analysis Aptitude |
| Work | Graduates, trainees, technicians, team leaders and supervisors. | |
| Operational | Operational roles in manufacturing, engineering, construction and transport. | Swift Comprehension Aptitude |
| Commercial | Commercial roles in sales, marketing and financial services. | |
| Customer | Customer interface roles in call centres, hospitality, leisure and health. | |
| Administrative | Administrative roles in private and public sector offices. | |
| Practical | Production, construction, engineering and scientific roles. | Swift Technical Aptitude |

# Analysis Aptitude Range

savilleconsulting



# Swift Combination Assessments

savilleconsulting

**Swift Analysis Aptitude (Invited Access):**

Unsupervised online 'Combination Assessment' at Professional / Work Aptitudes difficulty level:

- 8 Verbal Analysis items in 6 minutes
- 8 Numerical Analysis items in 6 minutes
- 8 Diagrammatic Analysis items in 6 minutes

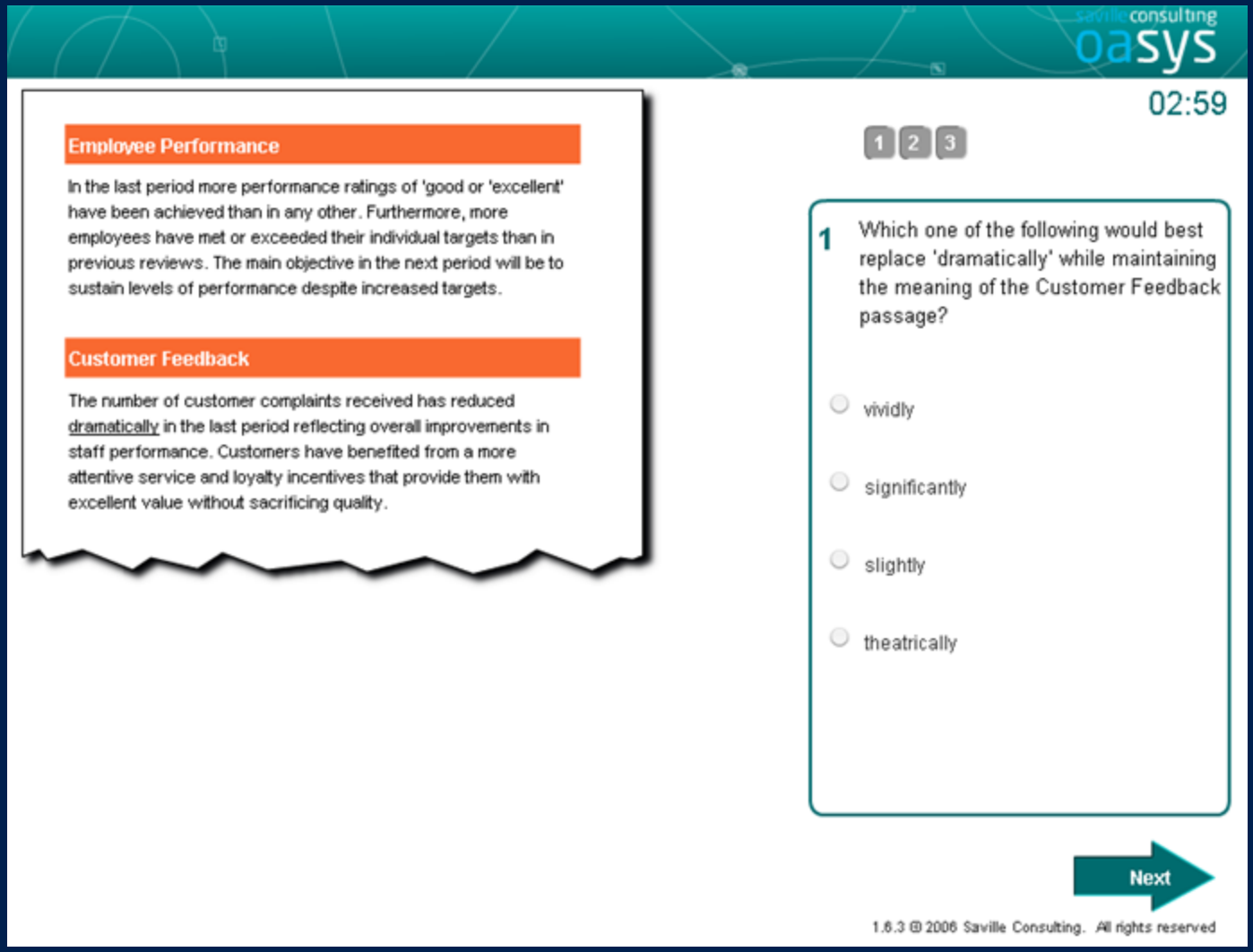
**Swift Analysis Aptitude (Supervised Access or Hard Copy):**

Supervised online or hard copy 'Combination Assessment' at Professional / Work Aptitudes difficulty level:

- 8 Verbal Analysis items in 6 minutes
- 8 Numerical Analysis items in 6 minutes
- 8 Diagrammatic Analysis items in 6 minutes

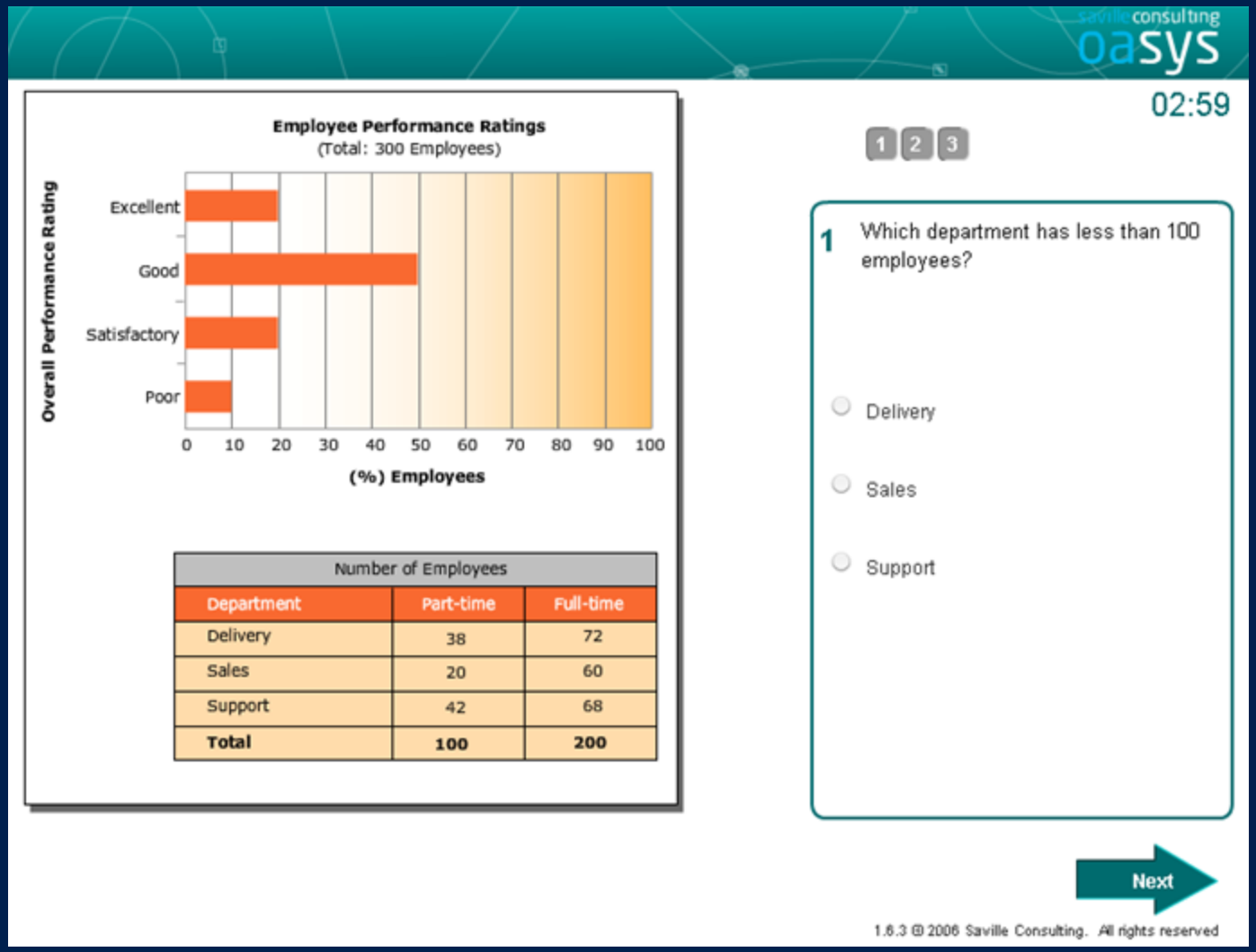
# Verbal Analysis Online

savilleconsulting



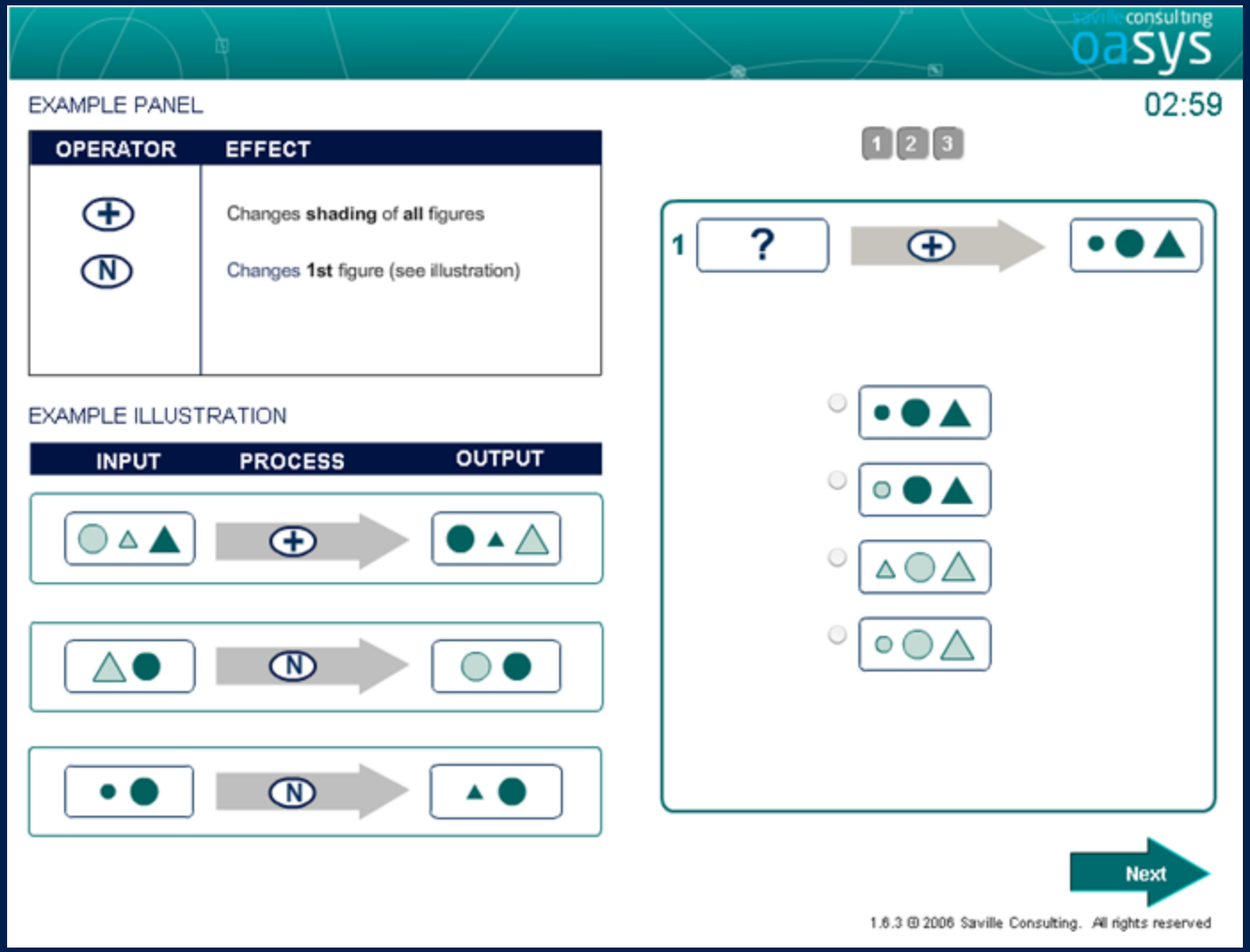
# Numerical Analysis Online

savilleconsulting



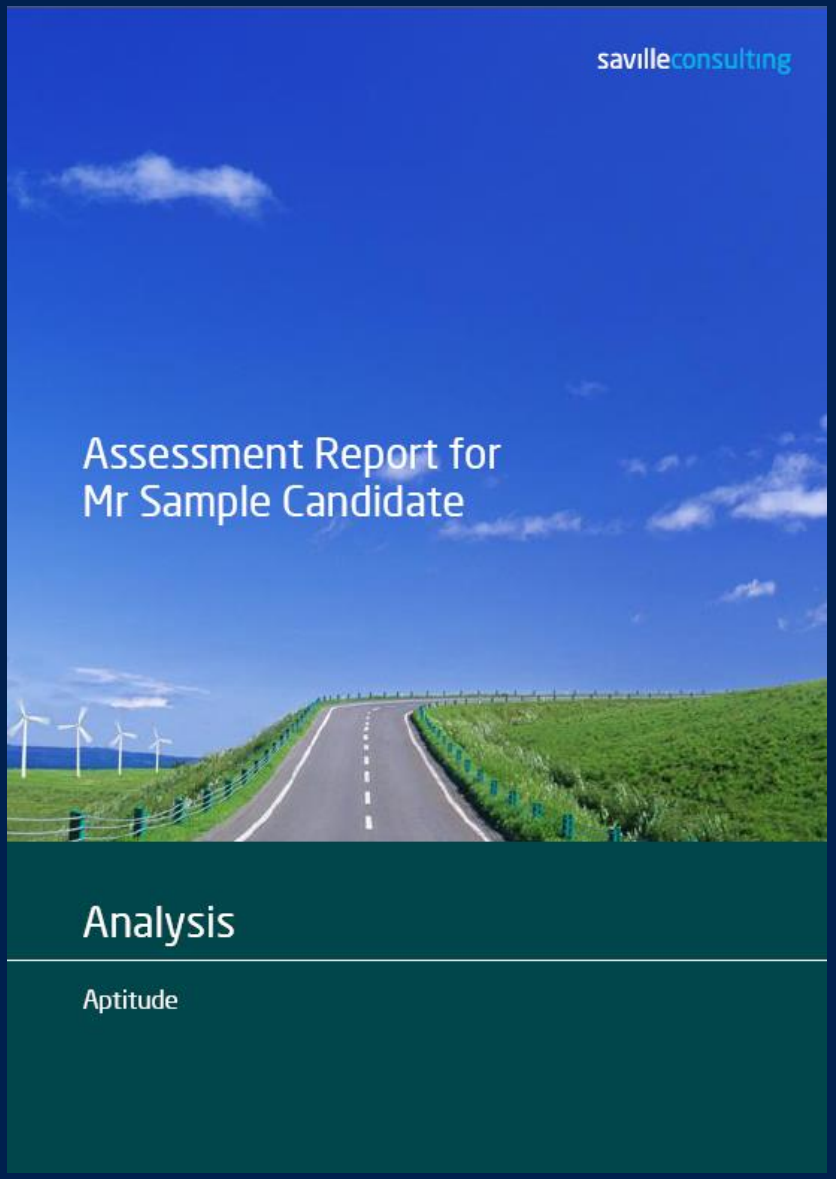
# Diagrammatic Analysis Online

savilleconsulting



# Assessment Report .pdf

savilleconsulting



# Designed for Test User, Candidate & Line Manager

savilleconsulting

## Introduction to Assessment Report

This report provides feedback on Sample Candidate's responses to the Analysis Aptitude assessment.

### Analysis Aptitude Profile

The assessment consists of three short tests measuring verbal, numerical and diagrammatic analysis aptitude areas that are important in the world of work for a variety of roles. The Analysis Aptitude Profile provides a summary of total and test taking style scores across the whole assessment, as well as sub-scores on the three aptitude areas covered in relation to the comparison group: Professionals & Graduates (IA; 2005)

### Total Score

The Total Score is the sum of correct answers across the verbal, numerical and diagrammatic analysis tests. It shows how well Sample Candidate has performed overall on the assessment.

### Test Taking Style

These scores indicate how quickly and accurately Sample Candidate completed the entire assessment.

**Accuracy:** concerns the proportion of answers that were correct.
**Speed:** concerns the number of questions answered.
**Caution:** is the difference between the Accuracy and Speed scores.

### Aptitude Area Sub-scores

These sub-scores provide information on how Sample Candidate performed on each of the three aptitude tests. The pattern of results indicates relative strengths and weaknesses across the following areas of aptitude:

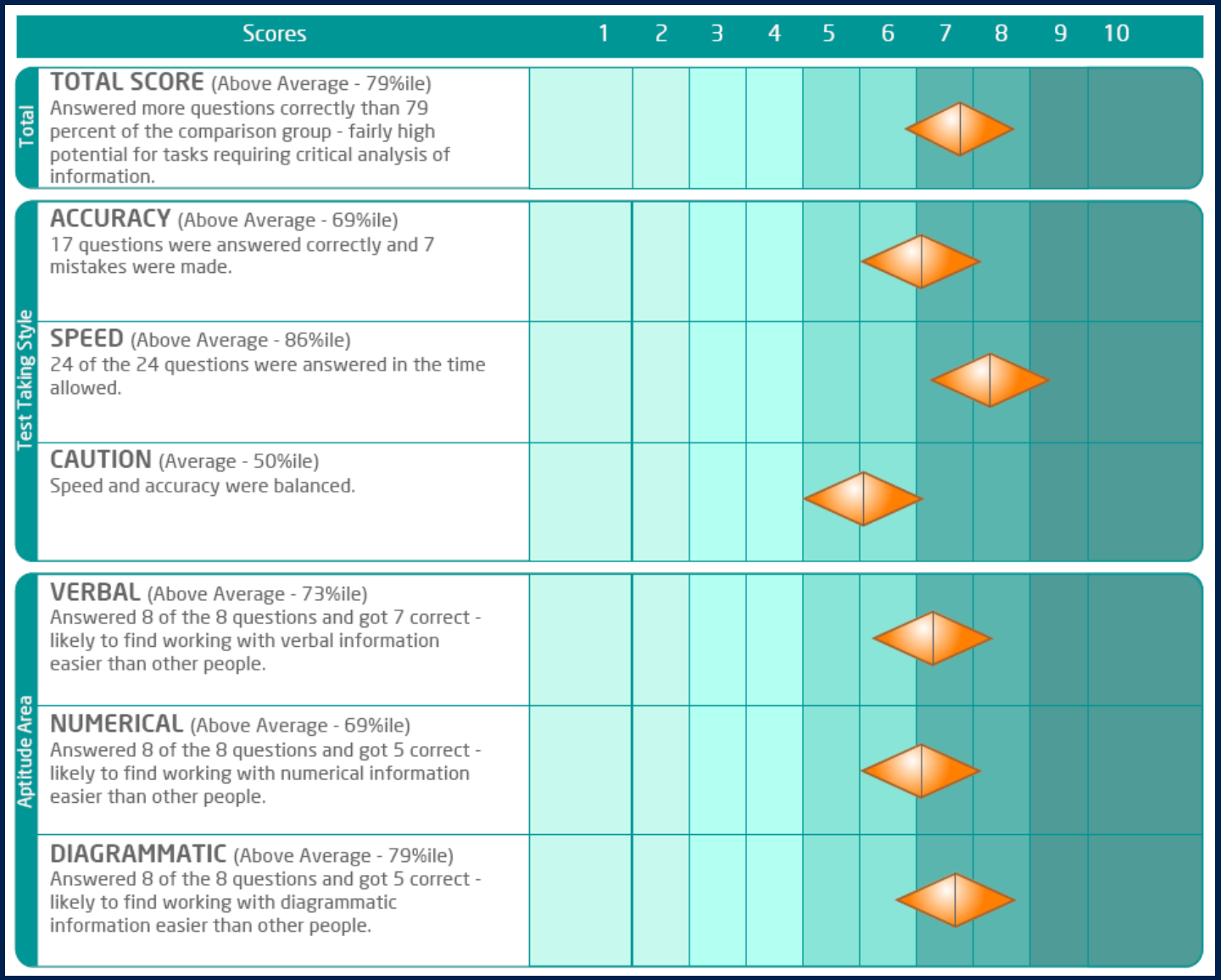
**Verbal -** assesses the ability to understand, interpret and evaluate written information, which is critical to success in areas such as Management, Law, Research, Sales and Administration.

**Numerical -** assesses the ability to understand, interpret and evaluate data, which is critical to success in areas such as Management, Finance, Engineering, Research, Sales and Administration.

**Diagrammatic -** assesses the ability to analyse diagrams, sequences and transformations, which is critical to success in areas such as Computer Programming, Engineering, Electronics and Science.

# Profile Chart with Test Taking Style & Item Type Sub-scores

savilleconsulting



# Swift Analysis Aptitude (HC & SA)

savilleconsulting

## Comparison Group: Professionals & Graduates (N=639)

| | Mean | SD | r Internal Consistency Reliability | r HC vs. IA Reliability[1] |
|---|---|---|---|---|
| **Total** | **15.42** | **4.38** | **.78** | **.72** |
| Accuracy | .70 | .17 | N/A | .66 |
| Speed | .91 | .11 | N/A | .38 |
| Caution | -.21 | .20 | N/A | .47 |
| Verbal | 5.33 | 1.83 | .56 | .51 |
| Numerical | 4.39 | 2.10 | .69 | .59 |
| Diagrammatic | 5.70 | 1.71 | .61 | .36 |

[1] Reliability Study (Alternate Form & Tests-Retest across administration mode with WWW IA followed by Hard Copy) on N=181 Research Officer Applicants; r=.69 before removal of 2 suspected 'Cheats' with 5 and 6 Sten performance drop

# Swift Analysis Aptitude (IA)

savilleconsulting

## Comparison Group: Professionals & Graduates (N=6745)

| | Mean | SD | r Internal Consistency Reliability | r HC vs. IA Concordance Reliability[1] |
|---|---|---|---|---|
| **Total** | **14.04** | **4.28** | **.76** | **.72** |
| Accuracy | .64 | .17 | N/A | .66 |
| Speed | .92 | .09 | N/A | .38 |
| Caution | -.28 | .18 | N/A | .47 |
| Verbal | 5.35 | 1.80 | .56 | .51 |
| Numerical | 4.54 | 2.08 | .70 | .59 |
| Diagrammatic | 4.15 | 1.72 | .51 | .36 |

[1] Reliability Study (Alternate Form & Tests-Retest across administration mode with WWW IA followed by Hard Copy) on N=181 Research Officer Applicants; r=.69 before removal of 2 suspected 'Cheats' with 5 and 6 Sten performance drop

# Swift Analysis Validation Study

savilleconsulting

Global Engineering Company Senior Managers (N=50) completed Swift Analysis 'Invited Access' online as part of Management Audit - Boss completed 'Work Performance Evaluation'.

Validities (unadjusted) against 'Overall Job Proficiency':

- Total Score .26*
- Accuracy .24*
- Speed .14
- Caution .11
- Verbal .04
- Numerical .21
- Diagrammatic .28*

# *Concordance Case Study*



This study is unusual as there is no restriction of range as individuals were told that online results would be used for decision-making and that supervised retesting would take place. For research purposes however all online participants were also invited to the hard copy session.

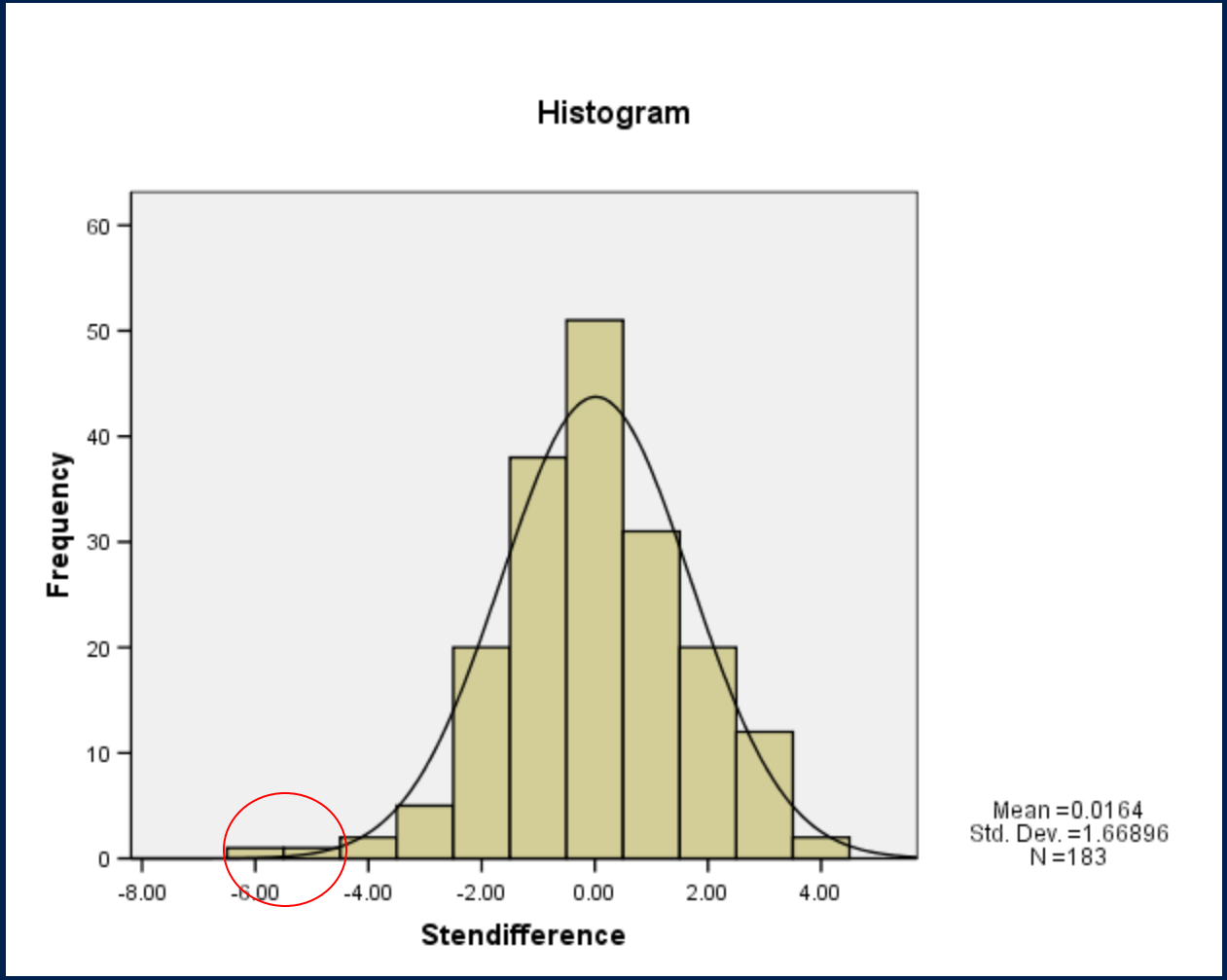
How do Sten Score differences across two occasions distribute simulating:

- 30%ile cut-off
- 70%ile cut-off
- How can we safeguard against 'Regression to the Mean'?

# SAA (IA vs. SA) Concordance Study

savilleconsulting

## Unsymmetrical Distribution identified 2 Suspected 'Cheats'



# Regression to the Mean 30%ile

savilleconsulting

## Regression to the Mean on SA Retest
## when using 30%ile Cut-off to screen out accounts for only ¼ Sten Drop

**Sten difference**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | -6.00 | 1 | .5 | .7 | .7 |
| | -5.00 | 1 | .5 | .7 | 1.5 |
| | -4.00 | 2 | .9 | 1.5 | 3.0 |
| | -3.00 | 5 | 2.4 | 3.7 | 6.7 |
| | -2.00 | 19 | 9.0 | 14.1 | 20.7 |
| | -1.00 | 29 | 13.7 | 21.5 | 42.2 |
| | .00 | 37 | 17.5 | 27.4 | 69.6 |
| | 1.00 | 23 | 10.8 | 17.0 | 86.7 |
| | 2.00 | 12 | 5.7 | 8.9 | 95.6 |
| | 3.00 | 4 | 1.9 | 3.0 | 98.5 |
| | 4.00 | 2 | .9 | 1.5 | 100.0 |
| | Total | 135 | 63.7 | 100.0 | |
| Missing | System | 77 | 36.3 | | |
| Total | | 212 | 100.0 | | |

# Regression to the Mean 70%ile

savilleconsulting

## Regression to the Mean on SA Retest when using 70%ile Cut-off to screen out accounts for 1 Sten Drop!

**Stendifference**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | -6.00 | 1 | 1.4 | 2.0 | 2.0 |
| | -5.00 | 1 | 1.4 | 2.0 | 4.0 |
| | -4.00 | 2 | 2.7 | 4.0 | 8.0 |
| | -3.00 | 2 | 2.7 | 4.0 | 12.0 |
| | -2.00 | 7 | 9.5 | 14.0 | 26.0 |
| | -1.00 | 17 | 23.0 | 34.0 | 60.0 |
| | .00 | 13 | 17.6 | 26.0 | 86.0 |
| | 1.00 | 6 | 8.1 | 12.0 | 98.0 |
| | 2.00 | 1 | 1.4 | 2.0 | 100.0 |
| | Total | 50 | 67.6 | 100.0 | |
| Missing | System | 24 | 32.4 | | |
| Total | | 74 | 100.0 | | |

# *Concordance Case Study*

savilleconsulting

| | | | Unsupervised | | Supervised | |
|---|---|---|---|---|---|---|
| | N | r | Mean | SD | Mean | SD |
| Unselected Group | 181 | .72 | 13.80 | 4.05 | 13.96 | 4.09 |
| Selected Group (>70%ile) | 56 | .50 | 18.50 | 1.85 | 17.38 | 2.78 |

Selecting only the top 30% of applicants (raw score of 16+) results in an average drop in scores of one Sten on the second testing occasion due to 'Regression to the Mean'.

'Restriction of Range' reduces the convergence correlation from .72 to .50.

The combined effect of 'Regression to the Mean' and high cut-off results in the majority of scores differing and true concordance being underestimated. The net effect will be that Majority of candidate scores will drop and it will 'appear' as if many candidates have cheated.

# *Concordance Case Study*

savilleconsulting

How do Sten Score differences across two occasions distribute:

- 30%ile cut-off – expect ¼ Sten 'Regression to the Mean'
- 70%ile cut-off – expect 1 Sten 'Regression to the Mean'
- We cannot safeguard against 'Regression to the Mean'!

# Discussion



- Even without any 'cheating' we can expect a majority of candidates to have substantially different scores

- The average standard score will drop on the supervised assessment where cutoff's are applied on the unsupervised.

- If we expect such differences WITHOUT CHEATING – what can we do in practice when we receive two very different scores?

  - Option 1: Verification Retesting
  - Option 2: Use Unsupervised & Supervised Data Independently

# Verification Retesting

savilleconsulting

- Score change of a particular size highlight potential cheats (e.g. 'not verified') - could we really safely eliminate these individuals from process???
- Potential need to Re-Retest
- Over-emphasis on verifying unsupervised (i.e. less trustworthy) data
- Issue of score increases – low scorers regress upwards to the mean
- Misleading documentation
- Testing may become discredited

# Use Unsupervised & Supervised Data Independently

savilleconsulting

- Use unsupervised (e.g. 'Invited Access') data for 'Screening Out'
- Use supervised data for 'Selecting In'
- DO NOT ATTEMPT INTERPRETATION OF DIFFERENCE SCORES AT THE INDIVIDUAL LEVEL
- MONITOR DIFFERENCE SCORES AT GROUP LEVEL
- No need to calculate difference scores i.e. hard copy ok
- No need for 'Re-Retest'
- Emphasis on supervised (i.e. more trustworthy) data
- Co-standardised or matched norm groups help
- Option of using more specific tests of key abilities

# Conclusion

savilleconsulting

- How many candidates will be labelled in Option 1 'Verification Testing' as 'Cheats' due to 'false positives'?
- How many test users will be confused by non-transparent reporting?
- How many line managers will 'switch off' testing because of cumbersome and erroneous procedures?
- We favour 'Option 2' – the independent use of data:
  - Simplicity
  - Higher weight on supervised test
  - Avoids ethical issues
  - Avoids 'cheating candidates' that did not cheat!