2019

MindmetriQ

Technical Fact Sheet

v3.0

MindmetriQ

1 Introduction

This technical fact sheet is intended to act as a **summary** of the research described further in the full MindmetriQ Technical Manual. The headline psychometric properties which this fact sheet reports on are:

Section 2: Construct Validity (**page 2**);
Section 3: Criterion-Related Validity (**page 9**);
Section 4: Group Differences (**page 11**); and
Section 5: Reliability (**page 14**)

References are included at the end of this fact sheet, and a brief overview of the MindmetriQ gamified assessments is given in Appendix A. For the full research we strongly recommend reading the full MindmetriQ Technical Manual.

2 Construct Validity

2.1 Test Partnership Insights

The Test Partnership Insights series is a selection of traditional validated aptitude tests that when combined, provide a measure of general cognitive ability (Schwencke & Guy, 2017). There are three individual aptitude tests within the Insights series:

Insights Verbal Reasoning (VR): This test is a measure of both understanding a passage of information, as well as what conclusions or assumptions can be drawn from it.

Insights Numerical Reasoning (NR): This assesses the ability to interpret and comprehend numerical information, in order to complete the required calculations.

Insights Inductive Reasoning (IR): This assesses the ability to recognise patterns and sequences in order to determine the next item in the sequence.

Table 2.1.1 presents correlations between the tests available within the Insights series and each Game-Based Assessment (GBA) of corresponding aptitude (i.e. numerical, verbal or inductive reasoning) within the MindmetriQ series: Net the Numbers (NTN), Link Swipe (LS), Pipe Puzzle (PP), Number Racer (NR), Word Logic (WL) and Shape Spinner (SS).

Insights Suite and Individual GBA Score Correlations			
GBA	Insights NR	Insights VR	Insights IR
NTN	0.65***		
NR	0.65***		
LS		0.71***	
WL		0.65***	
PP			0.58***
SS			0.44***

Note: * = Significant at p<0.05, ** = Significant at p<0.01, *** = Significant at p<0.001.

Table 2.1.1: Correlations between Individual MindmetriQ Scores and The Insights Suite.

The correlations have been corrected for reliability, but not restriction of range. All correlations are statistically significant at the p<.001 level, and can be interpreted as ranging from moderate to very large (Cohen, 1988).

Table 2.1.2 presents correlations for the following GBA combinations: Net the Numbers and Number Racer (Numerical), Link Swipe and Word Logic (Verbal), Pipe Puzzle and Shape Spinner (Logical), and all six of the GBAs within the MindmetriQ series (Full Battery).

Insights Suite and Combined GBA Score Correlations

GBA	Insights NR	Insights VR	Insights IR	Insights Full
Numerical	0.54***	0.18*	0.61***	0.56***
Verbal	0.60***	0.71***	0.52***	0.75***
Logical	0.53***	0.39***	0.72***	0.69***
Full Battery	0.57***	0.39***	0.67***	0.69***

*Note: * = Significant at p<0.05, ** = Significant at p<0.01, *** = Significant at p<0.001.*

Table 2.1.1: Correlations between Individual MindmetriQ Scores and The Insights Suite (n = 703).

All correlations have been corrected for reliability, but not restriction of range. Excluding the correlation between GBA Numerical and Insights VR, all combined score correlations can be interpreted as ranging from moderate to very large (Cohen, 1988). All correlations are statistically significant at the p<.001 level, excluding Insights VR and Numerical which is significant at the p<.05 level.

2.2 International Cognitive Ability Resource (ICAR)

The International Cognitive Ability Resource (ICAR) is an open source cognitive ability test (Condon & Revelle, 2014). The ICAR contains four subtests:

- *Verbal Reasoning (VR)*: The VR subtest measures a person's ability to work with words and sentences.
- *Letter Number Series (LNS)*: The LNS subtest measures a person's ability to identify patterns in alphanumerical data.
- *3D Rotation (3DR)*: The 3DR subtest measures a person's spatial reasoning ability by rotating three-dimensional shapes.
- *Matrix Reasoning (MR)*: The MR subtest measures a person's ability to solve problems and think logically.

The Full ICAR is a composite of all four subtests, which forms a measure of general cognitive ability.

Table 2.2.1 below displays the correlations between the ICAR and its subtests and each GBA within the MindmetriQ series: Net the Numbers (NTN), Link Swipe (LS), Pipe Puzzle (PP), Number Racer (NR), Word Logic (WL) and Shape Spinner (SS).

ICAR, ICAR Sub-Scales and Individual Score Correlations

GBA	ICAR VR	ICAR LNS	ICAR MR	ICAR 3DR	Full ICAR
NTN	0.80***	0.33***	0.84***	0.48***	0.79***
LS	0.89***	0.46***	0.18***	0.22***	0.63***
PP	0.48***	0.37***	0.58***	0.50***	0.61***
NR	0.93***	0.50***	0.82***	0.40***	0.88***
WL	0.81***	0.50***	0.40***	0.15**	0.66**
SS	0.63***	0.30***	0.38***	0.52***	0.57***

*Note: ^ = Not Significant, * = Significant at p<0.05, ** = Significant at p<0.01, *** = Significant at p<0.001.*

Table 2.2.1: Correlations between Individual MindmetriQ Scores and the ICAR (samples range from n = 223 to n = 374).

Correlations have been corrected for reliability and restriction of range. The correlations between the individual MindmetriQ scores and the Full ICAR can be interpreted as large (Cohen, 1988).

Table 2.2.2 on the next page presents correlations between the ICAR and the following GBA combinations: Net the Numbers and Number Racer (Numerical), Link Swipe and Word Logic (Verbal), Pipe Puzzle and Shape Spinner (Logical), and all six of the GBAs within the MindmetriQ series (Full Battery).

ICAR, ICAR Sub-Scales and Combined Score Correlations					
GBA	ICAR VR	ICAR LNS	ICAR MR	ICAR 3DR	Full ICAR
Numerical	0.86***	0.46***	0.88***	0.47***	0.89***
Verbal	0.91***	0.52***	0.31**	0.18*	0.69***
Logical	0.67***	0.43***	0.60***	0.60***	0.72***
Full Battery	0.71***	0.56***	0.76***	0.63***	0.81***

*Note: ^ = Not Significant, * = Significant at p<0.05, ** = Significant at p<0.01, *** = Significant at p<0.001.*

Table 2.2.2: Correlations between Combined MindmetriQ Scores and the ICAR (samples range from n = 223 to n = 374).

All correlations are statistically significant and have been corrected for reliability and for restriction of range. The correlations between the combined MindmetriQ scores and the Full ICAR can be interpreted as large (Cohen, 1988).

Although correlations for the subsets within the ICAR have been reported, it is important to note that only the correlations between the MindmetriQ tests and the Full ICAR are the most indicative, as the ICAR was intended to provide a measure of general cognitive ability for which the full scale is required.

2.3 Composite Cognitive Ability Tests (CCT)

Candidates who completed this construct validity measure were administered the following cognitive ability tests in succession of one another:

- *Cognitive Reflections Test (CRT)*: The CRT measures an individual's ability to reflect on problems, and override obvious (but incorrect) responses (Frederick, 2005).

- *BAROCO Short-form (BST)*: A short version of the original 100 item syllogism-solving problems (Shikishima, Yamagata, Hiraishi, Sugimoto, Murayama & Ando; 2011).

- *Berlin Numeracy Test (BNT)*: The BNT is a measure of numerical reasoning, risk literacy and statistical numeracy (Cokely, Galesic, Schulz, Ghazal & Garcia-Retamero; 2012).

- *Decision Making Competence Questionnaire* – Applying Decision Rules Sub-test (ADR): One sub-test of an original set of seven behavioural decision-making tasks (Bruine de bruin, Parker & Fischhoff; 2007).

Table 2.3.1 below displays the correlations between CCT and its subtests and each GBA within the MindmetriQ series: Net the Numbers (NTN), Link Swipe (LS), Pipe Puzzle (PP), Number Racer (NR), Word Logic (WL) and Shape Spinner (SS).

CCT, CCT Sub-Scales and Individual GBA Score Correlations					
GBA	CRT	BST	BNT	ADR	Full CCT
NTN	0.59***	0.46***	0.52**	0.77***	0.66***
LS	0.39***	0.39***	0.58***	0.62***	0.58***
PP	0.53***	0.51***	0.47***	0.85***	0.67***
NR	0.78***	0.35***	0.53***	0.86***	0.73***
WL	0.58***	0.71***	0.72***	0.86***	0.82***
SS	0.58***	0.61***	0.50***	0.62***	0.68***

*Note: ^ = Not Significant, * = Significant at p<0.05, ** = Significant at p<0.01, *** = Significant at p<0.001.*

Table 2.3.1: Correlations between Individual MindmetriQ Scores and CCT (samples range from n = 338 to n = 654)

Correlations in Table 2.3.1 have been corrected for reliability and for restriction of range. All correlations between the individual MindmetriQ scores and Full CCT can be interpreted as large (Cohen, 1988), and all correlations are significant at p<.001.

Table 2.3.2 below presents correlations between CCT measures and the following GBA combinations: Net the Numbers and Number Racer (Numerical), Link Swipe and Word Logic (Verbal), Pipe Puzzle and Shape Spinner (Logical), and all six of the GBAs within the MindmetriQ series (Full Battery).

CCT, CCT Sub-Scales and Combined GBA Score Correlations					
GBA	CRT	BST	BNT	ADR	Full CCT
Numerical	0.79***	0.47***	0.60***	0.96***	0.81***
Verbal	0.53***	0.59***	0.72***	0.80***	0.75***
Logical	0.70***	0.70***	0.62***	0.93***	0.82***
Full Battery	0.73***	0.62***	0.77***	0.95***	0.84***

*Note: ^ = Not Significant, * = Significant at p<0.05, ** = Significant at p<0.01, *** = Significant at p<0.001.*
Table 2.3.2: Correlations between Combined MindmetriQ Scores and CCT Measures (samples range from n = 251 to n = 654)

All correlations are statistically significant at p<.001 and they have been corrected for reliability and restriction of range. The correlations between the combined MindmetriQ scores and the Full CCT measure can be interpreted as large (Cohen, 1988).

2.4 Construct Validity Summary

These results suggest that the MindmetriQ series measures a similar underlying psychological construct to that measured by the ICAR, CCT measures, and by the Test Partnership Insights series. As all of these tests are validated measures of cognitive ability, this suggests evidence of construct validity for the MindmetriQ game-based assessments.

3 Criterion-Related Validity

3.1 Academic Achievement

Cognitive ability tests are useful predictors of many important life outcomes, in particular educational and occupational performance. Therefore, as cognitive ability measures, the MindmetriQ Series should display this kind of predictive validity.

A study was undertaken to determine the predictive validity of the MindmetriQ series with academic achievement, in particular GCSE results. Participants completed all six of the MindmetriQ GBAs along with a demographic questionnaire, in which participants provided their GCSE grades for the three compulsory GCSE subjects in the UK.

GCSE results were correlated with test scores from each of the MindmetriQ GBAs, these correlations can be seen in Table 3.1 below:

GCSE Performance and Individual GBA Score Correlations

GBA	GCSE Maths	GCSE Science	GCSE English	GCSE Overall
NTN	0.63***	0.53***	0.31***	0.57***
LS	0.27**	0.16^	0.29**	0.29**
SS	0.23**	0.13^	0.07^	0.18*
NR	0.45***	0.33***	0.10^	0.34***
WL	0.59***	0.67***	0.70***	0.73***
PP	0.32***	0.30***	0.09^	0.29**

*Note: ^ = Non Significant, * = Significant at p<0.05, ** = Significant at p<0.01, *** = Significant at p<0.001.*

Table 3.1: Correlations between Individual MindmetriQ Scores and GCSE Results (n = 125)

These figures have been corrected for reliability and for restriction of range. All correlations between individual MindmetriQ assessments and overall GCSE performance are statistically significant, and generally range from small to large in effect size (Cohen, 1988).

Table 3.2 below presents correlations for the following GBA combinations: Net the Numbers and Number Racer (Numerical), Link Swipe and Word Logic (Verbal), Pipe Puzzle and Shape Spinner (Logical), and all six of the GBAs within the MindmetriQ series (Full Battery).

GCSE Performance and Individual GBA Score Correlations				
GBA	GCSE Maths	GCSE Science	GCSE English	GCSE Overall
Numerical	0.53***	0.43***	0.20*	0.45***
Verbal	0.47***	0.50***	0.58***	0.58***
Logical	0.29**	0.31***	0.11^	0.32***
Full Battery	0.38***	0.43***	0.26**	0.46***

*Note: ^ = Non Significant, * = Significant at p<0.05, ** = Significant at p<0.01, *** = Significant at p<0.001.*

Table 3.2: Correlations between Combined MindmetriQ Scores and GCSE Results (n = 125)

These figures have been corrected for reliability and for restriction of range. All correlations between combined MindmetriQ scores and overall GCSE performance are statistically significant, and range from moderate to large in effect size (Cohen, 1988).

Note: Although ability tests are strong predictors of academic achievement, this does not imply that academic achievement is a strong predictor of job performance. Research shows job performance to be weakly-correlated with academic achievement, whereas ability tests are consistently shown to be the strongest single predictors of job performance.

4 Group Differences

4.1 Protected Groups

The following protected groups were investigated for evidence of average group differences and unfair discrimination:

- Gender: Participants that reported to be male were compared against participants reporting to be female.
- Ethnicity: Participants that reported being white were compared against those who reported to be BME participants.
- Age: Participants that reported being under the age of 30 were compared against the average scores of participants reporting to be over the age of 30.
- Sexual Orientation: Participants reporting not to identify as LGBTQIA+ were compared to participants that identified as being LGBTQIA+.
- Disability Status: Participants who do not have a disability were compared to those who consider themselves to have a disability.

Cohen's *d*					
GBA	Gender	Ethnicity	Age	Sexual Orientation	Disability Status
NTN	0.25	-0.08	0.20	-0.20	0.07
LS	-0.07	0.21	-0.04	0.15	0.13
PP	0.39	0.26	0.49	0.01	-0.01
NR	0.45	-0.01	0.52	0.11	0.04
WL	0.05	0.33	-0.15	0.17	-0.02
SS	-0.15	0.36	0.19	0.10	0.20

Table 4.1: Individual score effect sizes across different groups

Figures for group differences in terms of Cohen's d are presented in Table 4.1 for each GBA within the MindmetriQ series: Net the Numbers (NTN), Link Swipe (LS), Pipe Puzzle (PP), Number Racer (NR), Word Logic (WL) and Shape Spinner (SS).

All effect sizes (excluding "Age" for Number Racer) can be interpreted as being small or non-existent (Cohen, 1988). However, for Number Racer, the effect size for "Age" can be interpreted as being moderate (Cohen, 1988).

Some effect sizes are below the 0.5 threshold for practical importance, despite nearing this value. For example, "Age" for Pipes and "Gender" for Number Racer are 0.49 and 0.45, respectively. With these figures in mind, it is important for practitioners to consider their chosen combination of GBAs in terms of their candidate pool demographic.

However, as a general precaution to ensure that bias does not occur, we recommend that multiple GBAs from the MindmetriQ series are always used in combination when assessing candidates. For example, if an organisation wished to assess numerical reasoning, we would propose that they use both of the available numerical GBAs as opposed to a single GBA measure. This is useful as it would help to mitigate the effects of adverse impact in the case that it did arise, as well as providing a more holistic measure of the target variable.

Table 4.2 below presents group differences in terms of Cohen's d for the following GBA combinations: Net the Numbers and Number Racer (Numerical), Link Swipe and Word Logic (Verbal), Pipe Puzzle and Shape Spinner (SS), and all six of the GBAs within the MindmetriQ series (Full Battery).

Cohen's *d*					
GBA	Gender	Ethnicity	Age	Sexual Orientation	Disability Status
Numerical	0.35	-0.05	0.36	-0.05	0.06
Verbal	-0.01	0.27	-0.09	0.16	0.06
Logical	0.12	0.31	0.34	0.05	0.09
Full Battery	0.15	0.18	0.20	0.05	0.07

Table 4.2: Average score effect sizes across different groups

As presented in Table 4.2, we now see that all effect sizes for the combined GBA measures shown, can be interpreted as being small or non-existent (Cohen, 1988). Thus supporting our recommendation that multiple GBAs should be used during assessment.

4.2 Group Differences Summary

Group differences across the four studies generally report small to negligible group differences between focal and reference groups. The only effect size greater than 0.5 was observed in Number Racer when investigating Age. Therefore, we suggest that practitioners should take careful consideration when selecting which GBAs to administer, in reference to their candidate pool demographic. Preferably, we recommend that multiple GBAs from the MindmetriQ series are always used when assessing candidates, in order to provide a greater overview of candidates' ability and to serve as a precaution against the occurrence of adverse impact.

5 Reliability

5.1 Person Reliability

The Rasch model provides two separate forms of reliability: person reliability and item reliability. Person reliability is analogous to traditional notions of test reliability, indicating a level of precision when measuring the ability of the persons being measured. Rasch person reliability can be interpreted in the same frame of reference as Cronbach's Alpha, with a minimum of .70 as an acceptable level of test reliability.

All estimated person reliabilities based on the calibration samples are presented in Table 5.1 below:

MindmetriQ GBA	Number of Items	Approximate Test Time (including instructions)	Reliability
Net the Numbers	12	6 minutes	.71
Link Swipe	15	4 minutes	.75
Shape Spinner	10	4 minutes	.75
Number Racer	12	6 minutes	.73
Word Logic	12	7 minutes	.71
Pipe Puzzle	10	6 minutes	.84

Table 5.1: MindmetriQ Person Reliability Measures

5.2 Person Reliability Summary

All estimated reliabilities exceed the minimum required standard of .70, with estimated person reliabilities ranging from .71 to .84. This level of reliability ensures a high level of accuracy when assessing candidates, regardless of the MindmetriQ GBA used. Although these coefficients are already large, they are conservative values, because the method used to calculate the reliability underestimates reliability compared to Cronbach’s alpha.

5.3 Item Reliability

Rasch item reliability is based on the initial calibration samples used to generate item difficulty parameters. As the difficulty calibrations generated during item trialling are the difficulty parameters for the live assessment, the item reliabilities from initial calibration can be directly reported, rather than estimated. Item difficulty reliabilities for each MindmetriQ item bank are shown in Table 5.3 below.

MindmetriQ GBA	Sample Size	SD	S.E.	Reliability
Net the Numbers	3241	1.11	.12	.99
Link Swipe	6284	1.27	.16	.98
Shape Spinner	2661	.84	.10	.98
Number Racer	3473	1.22	.14	.99
Word Logic	2177	.83	.13	.97
Pipe Puzzle	1932	1.99	.18	.99

Table 5.3: MindmetriQ Item Reliability Measures

5.4 Reliability Summary

A general recommendation for Rasch item reliability is a minimum of .90. The observed item reliabilities range from .97 to .99, exceeding the minimum standard for Rasch item reliability. This high level of reliability ensures that item difficulty calibrations are accurate and precise, permitting their use in high stakes selection and assessment. It also ensures that item calibrations are highly stable, allowing these initial items to be used in common item linking, as part of ongoing trials for item bank expansion.

References

Bruine de Bruin, W., Parker, A. M., & Fischhoff, B. (2007). Individual differences in adult decision-making competence. Journal of personality and social psychology, 92(5), 938.

Cokely, E.T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin Numeracy Test. Judgment and Decision Making, 7, 25-47.

Cohen, J. (1988). Statistical power analysis for the behavioural sciences. 2nd.

Condon, D. M., & Revelle, W. (2014). The International Cognitive Ability Resource: Development and initial validation of a public-domain measure. Intelligence, 43, 52-64.

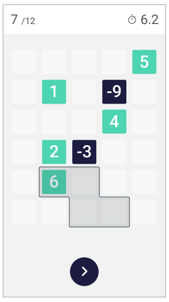
Frederick, S. (2005). Cognitive reflection and decision making. Journal of Economic Perspectives, 19(4), 25-42.

Schwencke, B., Guy, L. (2017). Test Partnership Insights Series: Technical Manual, London: UK

Shikishima, C., Yamagata, S., Hiraishi, K., Sugimoto, Y., Murayama, K., & Ando, J. (2011). A simple syllogism-solving test: Empirical findings and implications for g research. Intelligence, 39(2), 89-99.

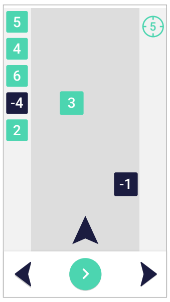
Appendices

APPENDIX A Overview of the MindmetriQ Gamified Assessment Series



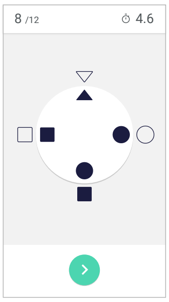
Net the Numbers (Numerical)

Facets measured: quantitative reasoning, working memory capacity, visual processing.



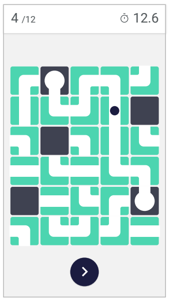
Number Racer (Numerical)

Facets measured: quantitative reasoning, perceptual speed, memory span.



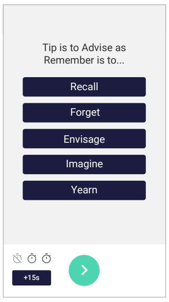
Shape Spinner (Logical)

Facets measured: speeded rotation, visualization, serial perceptual integration.



Pipe Puzzle (Logical)

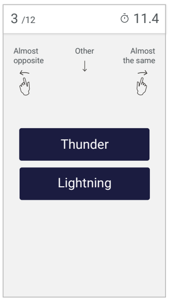
Facets measured: spatial scanning, visual memory, flexibility of closure.



Tip is to Advise as Remember is to...

Word Logic (Verbal)

Facets measured: reading decoding, processing verbal information, cloze reasoning (missing information).



Link Swipe (Verbal)

Facets measured: lexical knowledge, processing verbal information, grammatical sensitivity.