BEING AT FULL POTENTIAL · INNER DEVELOPMENT GOALS

How reliable and valid is the IDG assessment?

An honest test of whether the instrument measures a person's expression across the 25 IDG skills: what it does reliably, what it does not, and what we are changing as a result.

EXECUTIVE SUMMARY

The assessment measures something real and repeatable at the level of the whole profile, it detects development in proportion to how much development work was done, and its individual skill scores are noisier than the framework's structure implies, which is why the profile, not any single number, is what we ask coaches to work from.

Tested on 11,015 completed assessments from 10,417 people. The instrument as a whole is reliable, the five dimensions are reliable, and one organisation returned the same profile across three assessments more than two years apart. When the same 293 individuals were measured twice their scores moved beyond chance, and the size of the movement tracked how much inner-development work had actually been done.

Against that: only 5 of the 25 individual skill scales reach the conventional reliability threshold, so a single skill score for a single person carries little weight on its own. The five dimensions do separate, but **weakly**, and only once the questions they share are taken out of the comparison. Both findings are explained in full inside, including why the first is largely arithmetic.

0.91
Reliability of the instrument as a whole

5/25
Individual skills reaching the conventional threshold

0.93
Stability of one organisation's profile across three assessments

+1.30
Change in 293 people measured twice, beyond chance

22 September 2026
All figures reproducible from the platform data. Method at the back.

What we found, in one page

Six conclusions. Three are favourable, two are limitations we cannot yet resolve, and one is a change we are making.

1. **The instrument as a whole is reliable.** Cronbach's alpha of **0.91** across the 83 questions, and **0.81 to 0.89** for the five dimensions. Those are solid figures by any conventional standard.

2. **It returns the same reading from the same organisation over time.** One client assessed three times across 2,113 people: the profile of what stands out, measured against everyone else, reproduces at **r = 0.91 to 0.98**.

3. **And it moves when something changes.** 293 individuals matched across two administrations shifted **+1.30 points (t = 3.90)**, with all five dimensions and 18 of 25 skills moving beyond chance. Crucially, the size of the movement tracks how much inner-development work was actually done.

4. **Individual skill scores call for caution, not alarm.** Only **5 of the 25** skill scales reach the conventional 0.70 threshold, but **19 of the 20 that fall short are short rather than incoherent**; one is genuinely weak. The practical implication stands either way: **do not draw a conclusion about one person from one skill score.**

5. **The five dimensions separate, but only weakly, and only once you correct for the questions they share.** Compared directly they look identical. Restricted to skill pairs sharing no questions, same-dimension pairs correlate at **0.494 against 0.452** ($p = 0.008$). The structure is **supported, not proved**: one factor still explains **57%** of all variance.

6. **What we are changing.** Reduce question sharing so the skills can be measured separately, revisit three questions that do not behave like the scales they belong to, and report the overall and dimension scores with more confidence than individual skills.

How to read this paper

Every conclusion here is stated in plain English first. The statistics follow it, so that anyone who wants to check the work can. You do not need them to follow the argument.

Reliability, or "alpha"	Whether a set of questions hangs together. Imagine splitting the questions in half at random, scoring each half separately, and seeing whether the two halves rank people the same way. If they do, the scale is reliable. **Above 0.80 is good, above 0.70 acceptable, and below 0.60 means single scores should be read with caution.**
Correlation, or "r"	How closely two scores move together, from 0 (no relationship) to 1 (identical). Two skills correlating at 0.45 means people who score well on one tend to score well on the other, **but far from always**.
"Beyond chance", or "p"	Whether a result is bigger than the wobble you would expect from luck alone. We test it by shuffling the data thousands of times and seeing how often pure chance produces something as large. **p = 0.008 means about once in 125 shuffles**, so the real result is unlikely to be luck. Anything above p = 0.05, roughly one in twenty, we treat as noise.
Age-matched	Scores rise steadily with age, so any group that is younger or older than average will look different for that reason alone. Matching means comparing a group against a reference built to have **the same generational mix**, which removes age as an explanation.

> **One idea does most of the work in this paper.** A number can be wrong in two different ways: it can be *noisy*, wobbling around when nothing has really changed, or it can be *measuring the wrong thing*, steady but pointing somewhere else. Those are tested separately, and an instrument can pass one and fail the other. Keeping them apart is what separates a validation from a brochure.

Two different questions

"Is it validated?" bundles together two things that are tested differently and can come apart.

Reliability

Does it measure consistently?

If several questions are meant to capture the same quality, people should answer them consistently. A reliable scale gives a steady reading rather than a different answer each time you look.

Tested here with Cronbach's alpha, and with repeat administrations of the same instrument to the same organisation.

Validity

Does it measure the right thing?

A set of questions can be perfectly consistent and still measure something other than what it claims. Ten questions about shoe size would have excellent reliability and tell you nothing about presence.

Tested here through structure, through relationships the theory predicts, and through whether scores move when development happens.

> **Reliability is necessary but not sufficient.** This is the distinction most assessment marketing blurs. High alpha is often quoted as if it settled the question of validity. It does not, and we have separated them throughout.

The instrument

83 questions, answered on a four-point scale, producing scores on 25 inner skills grouped into five dimensions, plus an overall figure. The questions are deliberately **indirect**: rather than asking "how present are you?", which invites the answer the respondent thinks is wanted, they ask about behaviour and belief in specific situations and infer the quality from the pattern. That is a legitimate and common design, and it matters for how the results below should be read.

The data

Completed assessments	**11,015**
Distinct people	**10,417**
People with a complete 83-item response set, used for the reliability analysis	**3,258**
People assessed more than once	**514**
Of those, with at least 60 days between administrations	**179**

A completed assessment means all 83 questions answered. Partial assessments are excluded throughout: the platform scores an unanswered question as zero, which would enter an average as a very low score rather than as missing data.

Reliability: strong as a whole, weak in parts

SCALE	QUESTIONS	ALPHA
The whole instrument	83	**0.91**
Being	54	**0.89**
Thinking	37	**0.85**
Relating	30	**0.81**
Collaborating	30	**0.81**
Acting	41	**0.87**

Individual skills: **5 of 25** reach 0.70

By the usual convention 0.70 is acceptable and 0.80 good. The instrument as a whole and every dimension clear that comfortably. Most individual skills do not.

Why, and why it is less alarming than it looks

Alpha depends on two things: how much the questions in a scale agree with each other, and **how many questions there are**. At this instrument's typical level of agreement, the same questions of the same quality produce very different alphas depending only on count:

QUESTIONS IN SCALE	ALPHA IT CAN REACH
5	0.47
6	0.51
13	0.70
83	0.94

A six-question scale cannot reach 0.70 here however well written it is. Critical thinking rests on six questions. So a low skill alpha is substantially arithmetic rather than evidence of a badly built scale.

> **Short is not the same as faulty, and the two can be told apart.** How well a scale's questions agree with each other does not depend on how many there are. Measured that way, **19 of the 20 scales below 0.70 agree about as well as the rest of the instrument**: they are short, not incoherent. **Exactly one is genuinely weak**, where the questions agree at 0.073 against a median of 0.138 across the 25. That one is a defect, and it is on the list of what we are changing.
>
> **In plain terms.** A short scale is a short ruler: it still measures, just with wider gradations. Ten of these scales are five or six questions long, which is not enough to place one person precisely, though it is ample to compare two groups of people.
>
> **The practical rule that follows.** The overall score and the five dimension scores can carry weight. **A single skill score for a single person cannot, on its own.** The shape of a profile is more trustworthy than any one number in it. Our reports are written that way, and coaches should read them that way.

Which scales hold together, and which do not

Because alpha is distorted by scale length, the fairer comparison is the average agreement between questions within a scale, shown in the last column.

WEAKEST FIVE

SKILL	QS	ALPHA	AGREEMENT
Inclusive mindset and Intercultural competence	5	0.28	0.07
Co-creation skills	5	0.42	0.12
Empathy and Compassion	6	0.45	0.12
Perspective skills	6	0.47	0.13
Communication skills	6	0.47	0.13

STRONGEST FIVE

SKILL	QS	ALPHA	AGREEMENT
Openness and Learning Mindset	16	0.70	0.13
Conscious Use of Resources	10	0.70	0.19
Integrity and Authenticity	13	0.75	0.19
Long-term orientation and visioning	9	0.76	0.26
Resilience	12	0.77	0.22

Long-term orientation is a genuinely coherent scale: nine questions that agree with each other at 0.26. **Inclusive mindset** is weak on both counts: five questions that barely relate to one another at 0.07. Those are different problems and need different fixes.

Three questions that do not behave

Each question should correlate with the rest of its own scale. Three do not:

QUESTION	SCALE	CORRELATION WITH ITS OWN SCALE
"The harder we try the more likely we will succeed"	Presence	**0.01**
"When I disconnect from work I can access my creativity"	Presence	0.06
"What kind of training interests you most?"	Self-awareness	0.07

> **This is not a judgement on the wording or the intent.** The first question is reverse coded and the reasoning behind it is sound: trying harder is the opposite of slowing down and noticing, which is what presence asks for. The finding is narrower and purely empirical: across 3,258 people, answers to that question do not move with answers to the other ten questions in the same scale. Whatever it captures, it is not what they capture.

Validity: does the five-dimension structure hold?

It does, but weakly. The first test we ran said otherwise, and that test was confounded in a way that ran opposite to what we assumed.

The test that appears to say no

AVERAGE CORRELATION BETWEEN TWO SKILLS	R
In the **same** dimension	0.547
In **different** dimensions	0.544
Difference	**0.003**
Variance explained by a single underlying factor	57%

Read alone, that says the dimensions are decorative. But the 83 questions are **shared** between skills:

Across all 300 pairs of skills, correlation rises directly with how many questions the pair has in common: **0.461** with none, **0.568** with one, **0.664** with two, **0.748** with three or more.

In plain terms: many questions count towards more than one skill, so some pairs of skills are partly built from the same answers and are bound to look alike. To ask whether the dimensions are real, compare only skills built from entirely separate questions.

The test that can answer it

Restrict the comparison to the **156 pairs that share no questions at all**, where nothing mechanical forces them together:

PAIRS SHARING NO QUESTIONS	PAIRS	AVERAGE R
Same dimension	34	**0.494**
Different dimensions	122	0.452
Difference		**+0.042, beyond chance (p = 0.008)**

Not a scale-length artefact: the groups are balanced on questions per scale (8.2 against 8.3) and reliability (0.561 against 0.563).

The dimensions are real, and the overlap was hiding them. Shared questions are spread *across* dimensions more often than within one: **51% of cross-dimension pairs share questions against 32% of same-dimension pairs**. The overlap therefore inflates the cross-dimension side *more*, cancelling the genuine within-dimension excess and landing the headline difference on +0.003. We had assumed the confound ran the other way and read the null as "cannot tell". It was masking a real effect.

Stated at its strongest, and no further. The separation is small. Three of the five dimensions carry it clearly (Being, Acting, Relating); Thinking is flat and Collaborating runs slightly the other way, on 5 to 9 pairs each. All 25 skills remain strongly related to one another whichever way you cut it. **The framework's structure has support, not proof**, and a cleaner assignment of questions to skills is still what would settle it.

Validity: what the instrument does get right

1. It behaves as the theory predicts

If inner development genuinely accumulates with life experience, scores should rise with age. They do, monotonically, across four generations:

GENERATION	MEAN OVERALL	PEOPLE
Baby Boomers	**81.3**	150
Generation X	**78.8**	612
Millennials	**76.2**	799
Generation Z	**73.4**	1,470

A near eight-point climb from the youngest group to the oldest, in the right order, with no exceptions. This is **known-groups validity**: the instrument distinguishes groups that theory says should differ, in the direction theory predicts.

> **It also carries a practical warning.** Because the age effect is this large, comparing any group to a general average measures mostly its age composition. Every cohort comparison we publish is matched on generation for that reason, and any comparison that is not should be treated with suspicion, including comparisons of our own published averages.

2. Everyone's profile is uneven, and consistently so

The distance between a person's highest and lowest skill averages **26.7 points**, and **97% of people** exceed 15 points. An uneven profile is the normal condition rather than a sign of imbalance, which is worth saying to anyone reading their own report for the first time.

3. The same organisation gives the same reading

One client was assessed three times over more than two years, 2,113 people in total. Taking each wave's profile **relative to everyone else on the platform**, and comparing those shapes between waves:

COMPARISON	CORRELATION
Wave 1 with wave 2	**0.93**
Wave 1 with wave 3	**0.91**
Wave 2 with wave 3	**0.98**

The same skills stand out every time: long-term orientation about 4.5 points above the platform, empathy and compassion about 5 points below it, in all three administrations.

Measuring the profile *relative to the platform* matters. Presence is the lowest-scoring skill for everyone, so an organisation that keeps presence at the bottom is only reproducing the population pattern. Subtracting the platform average removes that, and what remains is specific to this organisation. It still reproduces.

Does it detect development?

Stability is only half an argument. An instrument that never moves is not stable, it is insensitive. So the question is whether scores change when something changes.

In plain terms: an instrument that never moves is not stable, it is blind. So the real test is not only whether it repeats, but whether it changes when the people being measured have changed.

The same people, measured twice

At the organisation described on the previous page, **293 individuals** could be matched across two administrations by employee number. This is the strongest design available to us: same people, same instrument, two points in time, with no question of who happened to take part.

MEASURE	CHANGE	T	BEYOND CHANCE?
Overall score	**+1.30**	3.90	**Yes**
All five dimensions	+1.16 to +1.64	2.97 to 4.45	Yes, all five
Individual skills moving beyond chance		18 of 25	about 1 expected

Replicated on a separate subset of 102 people measured at a third point: **+1.29, t = 2.26**.

An honest note on method. Compared at organisation level rather than person level, the same change reads **+0.51 and is indistinguishable from noise**, because the people taking part differed between waves. Pairing individuals removed that and revealed the effect. We report this because it shows how easily a real result can be missed, and because anyone checking our cohort-level figures should know they are the conservative ones.

The size of the change tracks the intervention

WHAT WAS DONE	CHANGE
Two years of individual development work (students, retakers)	**+6.6**
Comparable period, no targeted development	+3.0
Structural redesign aligning people to their strengths, no inner-development work	**+1.30**

This is the strongest evidence in the paper. A structural intervention that never targeted inner development produced about a fifth of the movement seen in dedicated individual development. The instrument does not simply move; it moves in proportion to how much of the relevant work was actually done. That pattern is difficult to produce by accident, and it is harder to explain away than any single reliability figure.

Stated at its strongest, and no further: three data points are not a dose-response curve, and none of these were controlled trials.

What this does not establish

The limitations that a careful reader would raise, stated before they have to.

- **It is a self-assessment.** Every score reflects how a person sees themselves. That is the intended use of the framework, and it is not an external measurement of how they behave. We have no data linking scores to observed behaviour or to outcomes, and until we do, no claim of that kind should be made on our behalf.
- **The 25 individual skills are still not shown to be distinct from each other.** The five dimensions separate (page 6); the 25 skills within them are a finer cut than the current question set can resolve. Anyone using individual skill scores as though they were independent measures is going beyond what we can support.
- **The change evidence rests on one organisation.** 293 paired individuals is a reasonable sample, but it is one client, one sector, one country. No second organisation on the platform has yet been assessed twice at a size that could confirm it, so the finding is **currently unreplicated**. That is the single biggest gap in this paper.
- **Retakers are self-selected.** People who take an assessment twice differ from those who take it once. The Windesheim comparison controls for this by comparing retakers with everyone else over the same period; the organisational data does not.
- **Repeat measurement is not neutral.** People who have seen the questions answer them differently, and growing self-awareness can lead someone to rate themselves *lower* because they see more clearly what a skill involves. A flat result after genuine development work is not necessarily a failure.
- **No test-retest interval was controlled.** Our repeat administrations happened when clients chose, not on a schedule designed to measure stability.
- **Six client datasets were bulk-imported** with a single timestamp for every participant, so for those cohorts we cannot establish when anyone answered. They are excluded from any time-based analysis here.

How large a repeat assessment has to be. Individual change is spread widely: the standard deviation of one person's movement between two assessments is **5.7 points**. Detecting a shift of +1.3 therefore needs roughly **75 people measured twice**; a shift of +3 needs about 14. **A repeat assessment of 20 people cannot settle whether an intervention worked**, whatever the numbers appear to say. Where a cohort is small, the honest reading is the direction and the shape of the profile, not a statistical verdict.

What would settle the open questions. A cleaner assignment of questions to skills would make the framework's structure testable. A study linking scores to behaviour rated by someone other than the respondent would establish criterion validity. Neither exists yet, and we would rather say so than imply otherwise.

What we are changing

1. **Reduce question sharing.** 83 questions currently produce **221 skill assignments**: each question feeds **2.7 skills on average**, 62 of the questions are used by more than one skill, one is used by 6, and only **17 belong to a single skill**. That overlap did not just obscure the dimension structure, it inverted the test for it. Moving towards one question, one skill needs either more questions or fewer separately reported skills, and that trade-off is what we are working through.

2. **Revisit the three questions on page 5** that do not move with their own scale, and the Inclusive mindset scale, which is the weakest of the 25 on both measures.

3. **Report confidence honestly in the product.** Overall and dimension scores carry weight; individual skill scores are indicative. Our reports already lead with the profile rather than with single skills, and we will make the distinction explicit.

4. **Build the evidence we do not have.** Replicate the change finding in other organisations, and design a study linking scores to something observed from outside the instrument.

5. **Republish these figures as the population grows.** Every number here is generated from the platform data by script rather than transcribed, so this document can be regenerated rather than rewritten.

> **Why publish a paper that criticises our own instrument.** Because the alternative is that someone else does it, less carefully, and because coaches deserve to know what the numbers they put in front of clients will and will not bear. An instrument whose limits are documented is more useful than one whose limits are unknown. We would rather be the people who tested it.

In one sentence

> The assessment measures something real and repeatable at the level of the whole profile, it detects development in proportion to how much development work was done, and its individual skill scores are noisier than the framework's structure implies, which is why the profile, not any single number, is what we ask coaches to work from.

Who did this analysis, and what that is worth

The analysis behind this paper was carried out by an **AI system** (Claude, made by Anthropic), working directly against the platform database under the direction of our founder. We are saying so for the same reason we published the unflattering findings: you should be able to judge the work by how it was done.

WHAT THAT ARRANGEMENT DOES BUY

Reproducibility. Every figure traces to a named script that can be re-run against the database. Nothing in this document was transcribed by hand, and it will be regenerated rather than rewritten as the population grows.

No stake in the answer. The analyst has no career, grant, citation record or reputation riding on whether the instrument performs well.

Patience for unglamorous checks. Twenty thousand reshuffles to test one number is cheap here, and that is how several findings were confirmed or thrown out.

WHAT IT DOES NOT BUY

Independence. The work was commissioned by the company that sells the instrument, run on that company's data, and reviewed by its founder before publication. An AI engaged by a vendor is not a disinterested third party and we are not going to claim it is.

Peer review, or pre-registration. What to test was decided as the work went along. That is how exploratory analysis proceeds, and it is also how results get shaped without anyone intending it.

Freedom from error. See below.

The mistakes made while producing this paper

FIRST CONCLUDED	WHY IT WAS WRONG	HOW IT WAS CAUGHT
The five dimensions do not separate; the data cannot settle it	The test was confounded by shared questions, in the **opposite** direction to the one assumed. Restricted to non-overlapping pairs, they do separate.	The founder asked whether the framing was more negative than the evidence warranted.
Individuals cannot be matched across repeat assessments	Stated three times from incomplete checks. An employee number was populated for every participant, and it is the basis of this paper's strongest finding.	The founder asked for it to be looked at again.
A client's three assessments formed one series	Only two of them did. Treating the third as part of the series produced an apparent decline that was partly an artefact.	The founder knew the client's history.

The pattern is the point. In every case the error was caught by a person who knew the business, not by the system that made it. So the honest description of what an AI analyst contributed here is **speed, consistency and reproducibility inside a process that still needed someone able to say "that does not sound right"**. Anyone claiming an AI is more objective than a human institution should look at this table first.

An invitation. We would rather this were checked than believed. If you work on measurement in this field and want to re-run any of it, we will help. The findings we would most like someone else to attack are the change result and the dimension structure.

Method

Data and scoring

All completed assessments on the platform as at 22 September 2026: **11,015 assessments from 10,417 people**. A completed assessment has all 83 questions answered. Where a person was assessed more than once, the most recent is used, except in the change analysis where both are. The same 83 questions underlie every product on the platform, with wording adapted per context.

Every figure is produced by the platform's own scoring engine, one assessment at a time, rather than reimplemented for this paper: the numbers here are the numbers each respondent sees in their own report.

Reliability

Cronbach's alpha computed on 3,258 people with a complete item set. Mean inter-item correlation derived from alpha and item count. Corrected item-total correlations exclude the item from its own scale total.

Structure

Correlations between the 25 skill scores, then an eigen-decomposition of that matrix to establish how much variance a single factor accounts for. Within- and between-dimension correlations compared both across all 300 skill pairs and restricted to the 156 pairs sharing no questions; the difference between those two groups is tested by 20,000 label permutations. Correcting for unreliability divides each correlation by the geometric mean of the two scales' alphas. The item-overlap figures are read from the scoring code.

Stability and change

One client assessed three times. Profile stability is the correlation between waves of each wave's deviation from the platform mean, skill by skill. The change analysis matches individuals across waves on employee number; the match was validated before use, with **392 of 392 cross-wave matches agreeing on gender** against roughly 60% expected from random pairing. Changes are paired t-tests. The detectable-change figures are the conventional 95% bound, using the observed spread of individual change.

What we excluded and why

- Partial assessments, because unanswered questions score zero and would depress averages.
- Six bulk-imported client datasets from time-based analysis, because every participant carries the same timestamp.
- Two further clients with repeat assessments, from the change analysis. Both were analysed and both returned nothing distinguishable from chance, but at 15 and 11 matched individuals neither could have detected an effect of the size observed here, so neither result carries information in either direction.

> **Reproducibility.** Every figure is generated from the platform data by script. No number in this document was transcribed by hand, and the document regenerates as the population grows. We will publish updated figures rather than restate these.