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. The overall score mostly reflects one shared quality, best read as a sense of meaning and purpose in life, work or studies, which accounts for about 91% of the differences in overall scores. 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. Factor analysis confirms that **one quality runs through all 83 questions**, whether they are asked about life, work or studies.

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; most of what each dimension score measures is that shared quality. 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

11 November 2026
All figures reproducible from the platform data. Method at the back.

Contents

The summary on page 3 carries the answer. The three parts that follow are ordered by verdict: what works, what calls for caution, and what we are improving.

BACKGROUND

What we found 3

How to read this paper 4

Two different questions 5

PART ONE · WHAT WORKS

The instrument as a whole is reliable 6

Validity: what the instrument does get right 7

Does it detect development? 8

Validity: does the five-dimension structure hold? 9

Interconnected, or indistinct? 10

The thread running through all 83 questions 11

What each score is made of 12

PART TWO · WHAT CALLS FOR CAUTION

Single skill scores are noisier than the framework implies 13

What this does not establish 14

Comparing generations, and how well the models fit 15

How large are the gaps between generations? 16

PART THREE · WHAT WE ARE IMPROVING

The scales and questions we need to fix 17

What we are changing 18

ABOUT THIS WORK

Who did this analysis, and what that is worth 19

Method 20

Method: factor analysis 21

What we found

Sorted three ways, which is also how the rest of this paper is ordered.

WHAT WORKS

1. **The instrument as a whole is reliable.** Alpha of **0.91** across the 83 questions and **0.81 to 0.89** for the five dimensions.

2. **It behaves as the theory predicts.** Scores rise with age across four generations, in the right order, with no exceptions.

3. **It returns the same reading from the same organisation.** Across three assessments and 2,113 people, the profile reproduces at **r = 0.91 to 0.98**.

4. **And it moves when something changes.** 293 people measured twice shifted **+1.30 points**, and the size of the shift tracked how much inner-development work was actually done.

5. **One quality runs through all 83 questions:** a sense of meaning and purpose, in life, work or studies. It accounts for about 91% of the differences in overall scores.

6. **The five dimensions do separate** once shared questions are taken out, though mostly they measure that same quality. Collaborating and Relating add the most of their own.

WHAT CALLS FOR CAUTION

7. **A single skill score for a single person carries little weight on its own.** Only **5 of 25** scales reach the usual threshold, though 19 of the 20 that fall short are short rather than incoherent. **Read the profile, not the number.**

8. **The change evidence rests on one organisation** and is currently unreplicated. A repeat assessment of fewer than about 75 people cannot detect a shift of the size we measured.

9. **Gaps between generations are approximate:** a few questions land differently with one age group.

WHAT WE ARE IMPROVING

10. **One scale does not hold together, and three questions do not behave** like the scales they belong to. Named in part three.

11. **Questions are shared between skills**, 2.7 per question on average, which is what made the dimension structure hard to test in the first place. Reducing that is the single most useful change available.

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 who took the IDG through its open door with all 83 answers recorded, used for the reliability and factor 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. The detailed analyses use the open IDG door only, so that every respondent took the same product in the same setting; about 11,200 people across all products have a full set of answers, and page 14 explains the narrower scope.

PART ONE · WHAT WORKS

The instrument as a whole is reliable

Reliability asks whether a set of questions hangs together. At the level of the whole instrument and of the five dimensions, it does, comfortably.

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**

By the usual convention 0.70 is acceptable and 0.80 good. **The whole instrument and every one of the five dimensions clear that comfortably**, on 3,258 people with a complete set of answers. These are solid figures by any conventional standard, and they are the scores our reports lead with.

> **What this earns.** An overall score and a dimension score can carry weight in a conversation. If someone scores low on Relating and high on Thinking, that difference is measured rather than noise, and it is a reasonable thing to build a coaching conversation on.
>
> Individual skill scores are a different matter, and that is taken up in part two.

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.
>
> Factor analysis confirms that each question works the same way at every age, and finds the size of the gaps between generations only approximate (page 15).

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.

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, and only three of the five dimensions carry it clearly. All 25 skills remain strongly related to one another whichever way you cut it, so **the structure has support, not proof**. **But "separate" may be the wrong test entirely, which the next page takes seriously. Pages 11 and 12 then name the shared part and measure what each dimension adds.**

Interconnected, or indistinct?

The previous page treats skills correlating across dimension boundaries as a measurement problem. There is a serious argument that it is not a problem at all.

The framework groups 25 capacities into five dimensions. **That grouping is a way of talking about inner development, not a claim that the psyche is built from five separable modules.** These capacities are deeply entangled: work on presence tends to surface in courage, and someone who becomes more self-aware rarely stays the same in how they relate to others. On that reading, capacities correlating across the boundaries is not a defect in the instrument. **It is the thing being measured.**

That reading is attractive, which is exactly why it needs testing. If the instrument were really measuring *one* quality over and over, the data would look a particular way. It does not. Five checks, each with what the opposite result would have meant:

1. **One broad quality does not explain everything.** The strongest single pattern running through all 25 skills accounts for **57%** of the variation between people. **43%** comes from somewhere else.

   Had the instrument been measuring one thing 25 times over, that figure would sit close to 100%.

2. **Four separate patterns are strong enough to count, not one.** The standard rule of thumb asks how many independent patterns are substantial enough to treat as real.

   One would mean everything collapses into a single trait wearing 25 different labels.

3. **Nobody's profile is flat.** The gap between a person's highest and lowest skill averages **26.7 points**, and **97%** of people differ by more than 15.

   If people were simply "high" or "low" across the board, these profiles would be level lines.

4. **That unevenness is reliable, not noise.** An organisation's profile shape reproduces at **r = 0.91 to 0.98** across three assessments spanning more than two years.

   A profile that reshuffled between assessments would mean the differences were random.

5. **When people develop, the capacities move together.** In the 293 individuals measured twice, **all five dimensions** shifted, not one or two.

   This speaks to the argument most directly: separable modules would show development concentrated where the work was done.

So there is a strong shared component *and* stable, reproducible differentiation on top of it. **That is the shape you would expect if inner capacities were interconnected but not identical**, which is what the framework has always claimed.

> **The guard on this argument.** An interpretation that can absorb any result is not an interpretation, it is an immunity. Each check above says what the opposite result would have meant, and two of them could plainly have gone against us: **development could have concentrated in one or two dimensions, and the overlap-free test on page 9 could have found no within-dimension excess.** Neither did, and both could have.

Page 11 names the shared component. The 57% above is a share of the 25 skill scores, each with its own part and its own noise; averaging 79 questions into the overall score cancels most of both, leaving the shared quality at about 91%. **None of this rescues the practical caution**, which is set out on page 13.

The thread running through all 83 questions

When the questions are left to show what they have in common, the answer is a sense of meaning and purpose: in life, in work, or in study.

Factor analysis asks a neutral question of the answers: which questions rise and fall together, and what do they share? It was not told what to look for. One broad pattern runs through most of the questions, and the questions that carry it most strongly are these:

STRENGTH	QUESTION (WORK WORDING; THE LIFE AND STUDENT WORDINGS BELOW)
0.73	I see the importance of my work Life: I see how what I do contributes to the bigger picture Studies: I see the importance of my studies
0.67	I have a role to play in meeting the growing challenges we face in society Life: … we face in the world
0.67	I clearly see the meaning in my work Life: I clearly see the meaning in my life Studies: I clearly see the meaning in my studies
0.67	Through my work I am making a positive difference in society Life: … in the world Studies: Through my studies I am being prepared to make a positive difference in society
0.61	When faced with a challenge, I pay attention to the lesson that it offers Life: same wording
0.60	When I take action it is usually with a sense of purpose and commitment Life: same wording

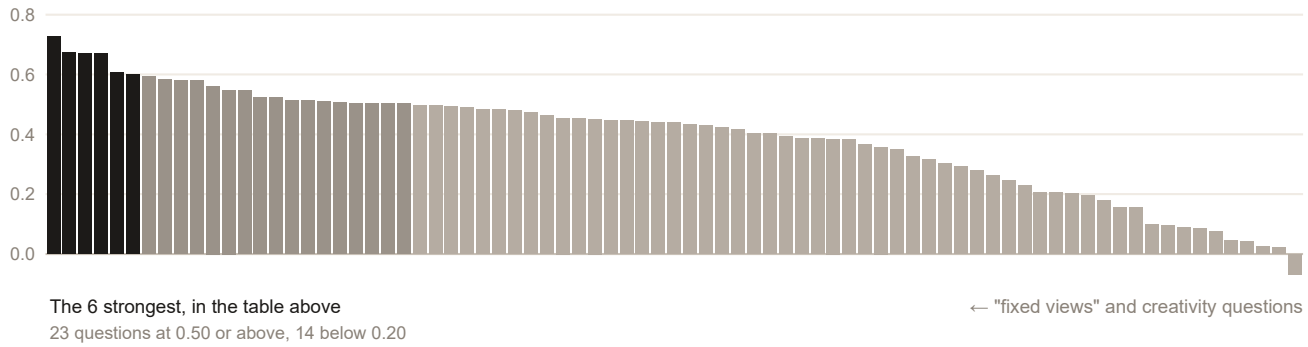

Figure 1. How strongly each of the 79 scored questions carries the shared pattern, strongest on the left (0 = not at all). Four of the 83 questions feed no skill score and are left out.

> **About 91% of the differences in overall scores come from this one quality.** That makes the overall score close to a single, coherent measure, which is what a headline score should be. Next to it sit growth (a clear idea how I want to grow, and taking the steps), excitement about the day ahead, pride in what one does, and clarity on one's own values.

This page names what page 10 found. "One strong shared pattern, with reliable differences on top" is still the shape of the data; the factor analysis shows what the shared pattern is. The differences on top are the subject of the next page.

What each score is made of

Every score is mostly that shared sense of meaning and purpose. Two of the five dimensions add something clearly their own.

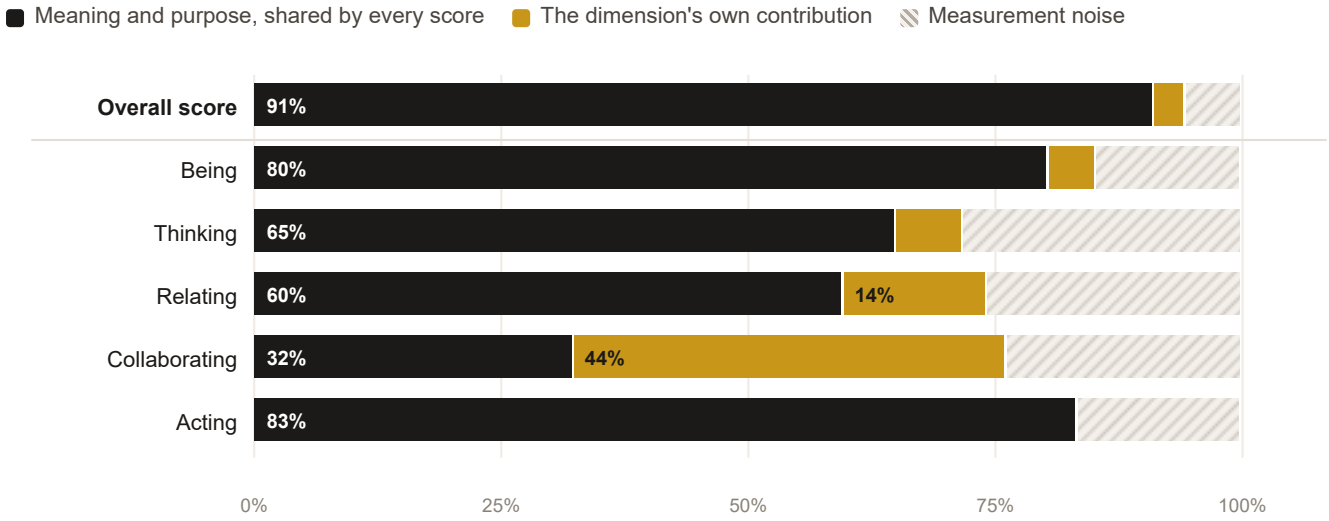

Figure 2. The share of each score that comes from the pattern shared by all questions (black), from the dimension itself (gold) and from measurement noise (hatched). Bifactor model, 1,629 people.

1. **Collaborating carries the most of its own**: 44% of its score, more than it shares with the rest. **Relating** adds 14%.

2. **Being and Acting are almost entirely the shared pattern.** They hold 42 of the 79 scored questions and lean most heavily on it, so the shared pattern is largely built from them.

   > *One reading: Being and Acting sit at the core of what the instrument measures. That is our interpretation; the numbers alone cannot show it.*

3. **Thinking adds little of its own** (7%), and its questions do not pull in one direction once the shared pattern is taken out.

How this sits with page 9. That page found the dimensions separating weakly, with Being and Acting separating best. This page finds Collaborating and Relating carrying the most of their own. The two tests ask different questions: whether skills in a dimension resemble each other (Being and Acting do, strongly), and what is left of a dimension once the shared pattern is removed (for Being and Acting, very little, because they largely *are* the shared pattern). Both are true. Part of Collaborating's own share may come from how its questions are worded, which page 15 explains.

PART TWO · WHAT CALLS FOR CAUTION

Single skill scores are noisier than the framework implies

Only **5 of the 25** individual skill scales reach 0.70. That sounds alarming and mostly is not, but it does change how one number should be read.

Most of the gap is arithmetic

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.

> **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**, and it is named in part three.
>
> **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.

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; 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.
- **Interconnection explains that limit; it does not remove it.** Chapter 8 argues that capacities moving together reflects the subject matter rather than a fault in the instrument, and the evidence supports it. **The practical consequence is unchanged either way**: a coach working from a single skill score is over-reading what the instrument can carry. Nor does interconnection excuse the question overlap, which is a measurement choice of ours and is still being reduced.
- **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 student 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.
- **One product, one setting.** These analyses use the open IDG door. Across all products, the structure fits client-organisation data noticeably less well, so we do not compare averages across products, and how far the findings carry over to organisations is not yet established.

> **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. A study giving the IDG alongside established measures of engagement and meaning in life would show what it adds to them. None of these exists yet, and we would rather say so than imply otherwise.

Comparing generations, and how well the models fit

The instrument works the same way for every generation. Comparing the average *level* of different generations calls for care.

Does a question mean the same thing at 22 and at 60?

We tested this in four steps, each stricter than the last, across Generation Z (1,470 people), Millennials (801) and Generation X with Baby Boomers (764):

HOLDS	HOLDS	HOLDS	PARTLY
The same overall pattern	**The answer options mean the same**	**Each question counts the same**	**The same level gives the same score**
baseline	change in fit -0.001	change in fit +0.018	change in fit -0.023

The first three steps hold: the generations read the answer options the same way, and each question carries the same weight for all of them. The strictest step holds only partly. On a few questions one generation answers higher or lower than its overall level would predict. A repeat in November 2026 on a larger group (4,061 people) gave the same result (strictest step -0.022, against -0.023 here). Page 16 names those questions and sizes the gaps.

> **What this changes.** The rise with age on page 7 keeps its direction, but the size of the gaps between generations should be read as approximate, and somewhat smaller than raw averages suggest: part of each gap comes from how particular questions land with each age group. Cohort comparisons are unaffected, because every comparison we publish is matched on generation, so both sides share the same age mix.

The strict step misses on one fit index (change -0.023 against the usual 0.010 limit) and passes easily on the other (+0.0016 against 0.015). Language and gender could not be tested: across all products only 144 of 9,560 people answered outside the default language, and gender is rarely recorded.

How well do the models fit?

Partly. On two standard measures the best model fits well (RMSEA 0.044, SRMR 0.056, both inside the usual limits). On a third it does not (CFI 0.80, where reviewers look for 0.90 or more). Long questionnaires with many indirect questions commonly show exactly this split: the broad structure is captured, many small question-to-question links are not.

Three things this analysis cannot settle

- **Questions where agreeing means less developed** ("People need to earn my trust", "When I win, usually someone else loses") form their own cluster. Such clusters often reflect how people answer reversed questions as much as what they believe. Most of them feed Collaborating, so part of its own share on page 12 may be wording.
- **The shared pattern resembles established measures of work engagement and meaning in life.** Whether the instrument measures something distinct from those needs a study that gives both to the same people. We would welcome partners for it.
- **The analysis shows that the skills move together with meaning and purpose.** It does not show that one causes the other.

How large are the gaps between generations?

Younger respondents express every dimension less strongly. Part of that gap comes from a handful of questions that depend on stage of life.

Allowing for how each question works, the November 2026 repeat (Generation Z 2,247, Millennials 970, Generation X with Baby Boomers 844) puts each generation on the same scale. Gaps are in standard deviations, against Generation X with Baby Boomers:

DIMENSION	GENERATION Z	MILLENNIALS
Being	-0.56	-0.28
Thinking	-0.65	-0.31
Relating	-0.57	-0.29
Collaborating	-0.91	-0.46
Acting	-0.78	-0.26

Because the strictest step on page 15 holds only partly, these sizes are approximate; the direction, and Collaborating showing the widest gap, are not in doubt. They are on the scale of the underlying factor, so they look larger than gaps in raw scores.

Which questions land differently

Twelve of the 79 questions account for 54% of the difference; the seven largest:

QUESTION (WORK WORDING)	GENERATION Z, AGAINST ITS GENERAL LEVEL	SHARE
I like things to be fully defined before I start working on them	Far below expected (agree much more often)	12%
Which statement best describes why you come to work?	Well below expected	7%
How do you acknowledge others?	Below expected	5%
Which statement best describes your wish for the future?	Slightly below expected	5%
The harder we try, the more likely we will succeed	Far below expected	5%
Saying “I don’t know” is a sign of strength	Well below expected (agree less often)	4%
What share of your talents is used in your current job?	Far below expected	4%

Several of these concern a current job or the reasons for working, which depend on stage of life as much as on inner development, and many younger respondents are students. Three questions run the other way, with younger respondents level with older ones: a team's shared goal, awareness of one's impact on others, and an employer encouraging self-exploration.

> **What follows.** Raw gaps between generations overstate the developmental gap somewhat. Re-estimating the gaps with these questions allowed to differ would give a corrected figure; that is the next step (page 18). Every cohort comparison we publish is already matched on generation, so none of them is affected.

PART THREE · WHAT WE ARE IMPROVING

The scales and questions we need to fix

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.

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 17** that do not move with their own scale, and the Inclusive mindset scale, which is the weakest of the 25 on both measures. And check whether the questions where agreeing means less developed act as a group of their own (page 15).

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, design a study linking scores to something observed from outside the instrument, and test how the IDG relates to established measures of engagement and meaning in life.

5. **Correct the generation gaps.** The questions that land differently by age are now named (page 16). Next, re-estimate the gaps with those questions allowed to differ, to give a corrected figure.

6. **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, best read as a sense of meaning and purpose expressed across 25 skills. 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 shared quality is meaning in work	The sample mixes work, life and student wordings: it is meaning and purpose in life, work or studies.	The founder pointed out the sample is not only work.

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. The factor analysis is described on page 21.

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.

Method: factor analysis

For readers who want to check pages 11, 12 and 15.

Sample and items. The 3,258 people with all 83 questions answered (as on page 6). Four questions feed no skill and are excluded, leaving 79. Answers are ordered categories (71 questions have four options), so all models use polychoric correlations and WLSMV estimation. The sample was split at random: exploratory analysis on one half, every confirmatory model on the other (1,629 each). Each question was assigned to the dimension it feeds most often; 40 ties were resolved on the exploratory half only. R 4.4.1, lavaan 0.7.3, psych 2.6.9, semTools 0.5.10.

Exploratory. First eigenvalues 15.5, 4.2, 3.2, 2.6, 2.1; the first is 3.7 times the second. Velicer's MAP suggests 7 factors, parallel analysis 16 (which over-extracts at this sample size). Five oblique factors group questions by theme across dimensions; 17 of 79 questions load below 0.30 on every factor. Schmid-Leiman ω_h on that solution: 0.629.

CONFIRMATORY MODEL	CFI	TLI	RMSEA	SRMR
One general factor	0.658	0.649	0.055	0.073
Five correlated dimensions	0.681	0.671	0.053	0.071
Five dimensions, each question on every dimension it feeds			not estimable (see text)	
Bifactor: general + five dimensions	0.748	0.735	0.048	0.065
Bifactor, each question on every dimension it feeds	0.798	0.779	0.044	0.056
Twenty-five skills, as scored			not estimable (see text)	

Scaled statistics. With questions feeding 2.7 skills on average, neither the full-overlap model nor the 25-skill model (221 assignments, two hours) converges, also with eight times the memory (November 2026).

BIFACTOR, GENERAL + FIVE DIMENSIONS	QUESTIONS	ω	ω_h GENERAL	ω_{hs} OWN
Being	26	0.85	0.80	**0.05**
Thinking	12	0.72	0.65	**0.07**
Relating	11	0.74	0.60	**0.14**
Collaborating	14	0.76	0.32	**0.44**
Acting	16	0.83	0.83	**0.00**
Overall score	79	0.94	**0.91**	

Explained common variance (ECV) 0.70. The other bifactor puts the same two on top (Collaborating 0.27, Relating 0.06).

INVARIANCE ACROSS GENERATIONS (FIVE-DIMENSION MODEL)	CFI	RMSEA	SRMR
The same overall pattern	0.738	0.046	0.071
The answer options mean the same	0.737	0.046	0.071
Each question counts the same	0.755	0.044	0.073
The same level gives the same score	0.732	0.046	0.073

Wu and Estabrook (2016) identification. CFI rising at the loadings step is a known artefact of scaled statistics. Latent means and the questions behind the partial result (page 16): November 2026 repeat, 4,061 respondents, score tests summed per question.