Becoming You Labs

RESEARCH REPORT

OCTOBER 2026

Feedback you can trust:

Development and validation of PIE360

How the revised PIE360 was built and tested, and what the evidence shows.

AUTHORS

Daniel B. Davis, PhD
Head of Research, Becoming You Labs

Daphna Harel, PhD
Professor of Applied Statistics, New York University

Cite as: Davis, D. B., & Harel, D. (2026). *Feedback you can trust: Development and validation of PIE360.* Becoming You Labs.

BECOMING YOU LABS

PIE360
OCTOBER 2026

Contents

	Executive summary	3
01	Introduction	4
02	Stage 1: What real-world use of the original form showed	10
03	Stage 2: Expert review and literature grounding	14
04	Stage 3: Statement sorts and web probing	17
05	Stage 4: Confirmatory factor analysis	21
06	What the results mean and what comes next	29
	References	32
	Appendix A. Literature crosswalk	35
	Appendix B. Technical details	37

Executive summary

PIE360 is a 360-degree feedback tool from Becoming You Labs that measures three domains of work behavior: People, Ideas, and Execution. This paper reports how the revised PIE360 was built and tested, and what the evidence shows.

36
Behaviors, 12 per domain

4
Development stages

405
Working U.S. adults in the national test

10
Rater minimum

WHAT PIE360 MEASURES

Thirty-six short items, 12 per domain, each describing one behavior a coworker can observe, from listening closely to meeting deadlines. Every item is grounded in published research on work performance, and no item assumes the person manages others.

People
How someone treats and affects others

Ideas
How someone generates and engages with ideas

Execution
How reliably someone delivers

HOW IT WAS BUILT

Four stages, each building on the last:

STAGE 1
An analysis of the original form as 669 people and their evaluators had used it.

STAGE 2
An expert panel that defined what to measure and checked it against seven established models of job performance.

STAGE 3
Statement sorts in which 39 readers tested whether the items were understood as intended, leading to clearer wording.

STAGE 4
A national test in which 405 working U.S. adults each rated a coworker.

WHAT THE EVIDENCE SHOWS

- The 36 items form the three intended domains, with good statistical fit and every item strongly tied to its domain.
- The items work the same way across every rater group tested, including rater sex, rater ethnicity, and the working relationship.
- Each domain is highly reliable.
- The new rating scale leaves more room at the top than the original form did.

WHAT IT MEANS FOR USERS

Read the three domains together, as a profile. Invite at least 10 raters, the minimum PIE360 requires and a number research supports. Use the results for development: the gap between how people see themselves and how others see them is often the best place to start.

SECTION 01

01

Introduction

1.1	1.2	1.3
Why 360-degree feedback is worth doing	What PIE360 offers	What this paper reports
P. 5	P. 5	P. 8

1.1

Why 360-degree feedback is worth doing

Most people hear about their work from one direction: their manager. A 360-degree assessment widens the view. The person rates themselves, and coworkers who see their work from different positions, such as managers, peers, and direct reports, rate the same behaviors. The ratings are pooled and returned to the person, most often to guide their development (Bracken, Rose, & Church, 2016).

Asking several people matters because different people see different things. In a review of published studies, Conway and Huffcutt (1997) found that supervisors, peers, and direct reports agreed only modestly when rating the same person (correlations of .22 to .34), and that much of any single rater's judgment reflects that rater's own perspective. Averaging several raters cancels much of that individual slant, which is why a 360 can show something a single review cannot. A 360 also lets people compare how they see themselves with how others see them, a comparison research links to how effective people are judged to be at work (Atwater, Ostroff, Yammarino, & Fleenor, 1998; Fleenor, Smither, Atwater, Braddy, & Sturm, 2010).

Feedback helps most when it is designed and used well. Across 24 studies that followed people over time, ratings improved after 360-degree feedback, though the average gains were small, and improvement was most likely when people set goals and acted on them (Smither, London, & Reilly, 2005). In one study, managers who worked with a coach after receiving feedback improved slightly more than those who did not (Smither, London, Flautt, Vargas, & Kucine, 2003). Poorly designed feedback can backfire: in a large review, feedback made performance worse in more than a third of cases (Kluger & DeNisi, 1996). The research points to relevant content, credible data, and real follow-through (Bracken & Rose, 2011; Nowack & Mashihi, 2012). For a tool like PIE360, that means careful item writing, a clear rating scale, enough raters, and a real conversation about the results.

1.2

What PIE360 offers

PIE360 is a 360-degree tool from Becoming You Labs, organized around three domains from the framework developed by Dr. Suzy Welch: People (how someone treats and affects others), Ideas (how someone generates and engages with ideas), and Execution (how reliably someone delivers). The revised form has 36 items, 12 in each domain (Table 1). Each item describes a single behavior a coworker can observe, such as "[NAME] does what they say they'll do," rated from 1 (Not at all) to 10 (Exceptionally well, one of the best I know). People answer the same items about themselves. No item assumes that the person manages others, so the tool suits individual contributors and people early in their careers as well as managers.

TABLE 1

The revised PIE360: 36 behaviors and the research behind them

ITEM		WHAT IT CAPTURES	RESEARCH BASIS
People			
P1	Authenticity	Coming across as genuine	Apparent sincerity (Ferris et al., 2005)
P2	Emotional Regulation	Staying composed when interactions get tense	Emotion regulation (Gross, 1998); handling work stress (Pulakos et al., 2000)
P3	Respect for Other Views	Respecting people who see things differently	Perspective taking (Parker & Axtell, 2001)
P4	Empathy	Recognizing how others feel	Empathy (Davis, 1983)
P5	Interpersonal Adaptability	Adjusting one's style to different people	Interpersonal adaptability (Pulakos et al., 2000); social astuteness (Ferris et al., 2005)
P6	Interpersonal Consistency	Treating people well regardless of one's mood	Emotional volatility, reversed (Soto & John, 2017)
P7	Active Listening	Giving others full attention	Active-empathic listening (Bodie, 2011)
P8	Positive Energy	Lifting the energy of the group	Relational energy (Owens et al., 2016)
P9	Constructive Disagreement	Keeping disagreement about the work, not the person	Task versus relationship conflict (Jehn, 1995; De Dreu & Weingart, 2003)
P10	Helpfulness	Helping coworkers who need it	Helping behavior (Van Dyne & LePine, 1998; Podsakoff et al., 2000)
P11	Recognizing Others	Giving others credit	Interpersonal facilitation (Van Scotter & Motowidlo, 1996); citizenship behavior (Podsakoff et al., 2000)
P12	Consideration for Others	Thinking about how one's actions affect others	Courtesy (Organ, 1988; Podsakoff et al., 2000)
Ideas			
I1	Idea Fluency	Generating many ideas	Ideational fluency (Guilford, 1967); idea generation (Janssen, 2000)
I2	Discussion Engagement	Engaging with ideas as they are discussed	Personal engagement at work (Kahn, 1990)
I3	Building on Ideas	Extending others' ideas usefully	Collective creativity (Hargadon & Bechky, 2006)

ITEM		WHAT IT CAPTURES	RESEARCH BASIS
I4	Discussion Readiness	Thinking a topic through before discussing it	Cognitive engagement (Rich, LePine, & Crawford, 2010)
I5	Originality	Coming up with novel ideas	Originality (Guilford, 1967); employee creativity (Zhou & George, 2001)
I6	Raising Concerns	Flagging possible problems with proposals	Prohibitive voice (Liang, Farh, & Farh, 2012)
I7	Problem Framing	Reframing problems to open new possibilities	Problem construction (Reiter-Palmon, Mumford, O'Connor Boes, & Runco, 1997)
I8	Idea Development	Refining ideas until they work	Innovative work behavior (Janssen, 2000; De Jong & Den Hartog, 2010)
I9	Flexible Thinking	Updating ideas in light of new information	Cognitive flexibility (Martin & Rubin, 1995); adapting and coping (Bartram, 2005)
I10	Continuous Learning	Continuing to learn about new ideas	Learning work tasks (Pulakos et al., 2000); learning goal orientation (VandeWalle, 1997)
I11	Analytical Thinking	Working through complex information	Analyzing and interpreting (Bartram, 2005); analyzing data or information (National Center for O*NET Development, n.d.)
I12	Sound Reasoning	Grounding ideas in clear reasons	Analyzing and interpreting (Bartram, 2005); evaluating arguments on their merits (Stanovich & West, 1997)
Execution			
E1	Keeping Your Word	Following through on what one says	Behavioral integrity (Simons, 2002)
E2	Honoring Commitments	Keeping commitments even at a cost	Responsibility (Soto & John, 2017); job dedication (Van Scotter & Motowidlo, 1996)
E3	Self-Management	Managing one's own responsibilities	Responsibility (Soto & John, 2017); job dedication (Van Scotter & Motowidlo, 1996)
E4	Task Completion	Seeing tasks through to the end	Perseverance of effort (Duckworth, Peterson, Matthews, & Kelly, 2007); job dedication (Van Scotter & Motowidlo, 1996)
E5	Timeliness	Meeting deadlines	Time management (Claessens, van Eerde, Rutte, & Roe, 2007); individual work performance (Koopmans et al., 2013)

ITEM		WHAT IT CAPTURES	RESEARCH BASIS
E6	Dependability	Following through for people who rely on you	Job dedication (Van Scotter & Motowidlo, 1996)
E7	Accountability	Owning results, good or bad	Responsibility (Soto & John, 2017)
E8	Persistence	Sticking with tedious work	Self-discipline (Costa & McCrae, 1992); demonstrating effort (Campbell, 1990)
E9	Prioritization	Doing the most important work first	Individual work performance (Koopmans et al., 2013); time management (Claessens et al., 2007)
E10	Thoroughness	Doing careful, high-quality work	Organizing and executing (Bartram, 2005); work quality (Koopmans et al., 2011)
E11	Following Procedures	Following agreed processes	Following organizational rules and procedures (Borman & Motowidlo, 1993); maintaining personal discipline (Campbell, 1990)
E12	Staying Organized	Keeping work in order	Organizing and executing (Bartram, 2005); individual work performance (Koopmans et al., 2013)

Note. Each item describes one observable behavior, rated from 1 (Not at all) to 10 (Exceptionally well, one of the best I know); the self-assessment uses the same items in the first person. Some research anchors match the behavior closely and others are related constructs (Section 3.2). Appendix A maps the items to seven broader models of job performance.

PIE360 is also shorter than many enterprise 360 tools, which matters because longer questionnaires draw fewer people to start and produce lower-quality answers toward the end (Galesic & Bosnjak, 2009). In testing, raters finished a 45-item version in a median of 4.5 minutes; the 36-item form should take about 3.5. Results are released only after at least 10 raters respond, for reasons Section 6 explains.

1.3

What this paper reports

This is the first research report on PIE360. It reports how the revised form was developed and tested in four stages, each building on the last (Figure 1):

1. **Real-world data.** An analysis of the original 28-item form as 669 people and their evaluators had used it.
2. **Expert review.** An expert panel defined what PIE360 should measure and grounded each quality in established research.
3. **Statement sorts.** Thirty-nine students sorted draft statements into domains and explained how they read them.
4. **A national test.** Working U.S. adults, 405 in all, each rated a coworker, and a confirmatory factor analysis tested whether the items formed the three intended domains.

Four stages took PIE360 from its original 28 items to a tested 36

STAGE 1	STAGE 2	STAGE 3	STAGE 4	RESULT
Real-world data Original 28-item form in use 669 people 8,340 ratings	**Expert review** 21 core constructs 7 job-performance frameworks 36 statements	**Statement sorts** 39 sorters 24 of 36 passed both sorts Wording refined	**National test** 405 coworker raters 45 items tested Best 36 selected	**Revised PIE360** 36 items, 12 per domain Good fit and high reliability

FIGURE 1 **How the revised PIE360 was built.**

Development stages, 2026 · numbers from Sections 2 to 5

SECTION 02

02

Stage 1: What real-world use of the original form showed

Dr. Harel analyzed the original 28-item PIE360 as people had actually used it. Three findings shaped the revision: scores needed many raters to be stable, three reverse-worded items did not fit, and ratings bunched near the top of the scale.

STAGE 1	STAGE 2	STAGE 3	STAGE 4
Real-world data	Expert review	Statement sorts	National test
P. 10	P. 14	P. 17	P. 21

2.1

Data and instrument

The data were every assessment completed on the original form, after test accounts and incomplete or invalid records were removed. We kept everyone with a completed self-assessment and ratings from at least three evaluators: 669 people and 8,340 evaluator ratings. The typical person had 12 evaluators.

The original form had 28 items (10 People, 9 Ideas, and 9 Execution), rated from 1 to 10 with the ends labeled "Occasionally" and "Regularly," so it asked how often a behavior occurred. Three items were reverse-worded, meaning a high rating signaled less of the quality being measured.

2.2

Item-level findings

A domain's items should move together if they measure the same thing. This internal consistency is measured with Cronbach's alpha, where .70 is usually treated as acceptable and .80 as good. On self-assessments, Execution was strong (.89), while People (.73) and Ideas (.70) sat at the low edge of acceptable (Table 2). The weakness traced largely to the three reverse-worded items, which did not move with the rest of their domains even after reverse scoring; setting them aside raised alpha to .80 for People and .78 for Ideas.

The expert panel's review (Stage 2) found related problems. Some items asked about two behaviors at once, some asked about feelings an evaluator cannot see, and some self and evaluator versions asked slightly different things. Reviews of 360 practice flag exactly these features as sources of rater confusion (Bracken, Timmreck, Fleenor, & Summers, 2001).

2.3

How many raters it took

The central question for any 360 is how much a person's score would change if a different set of raters had been asked. Single-rater reliability answers it: it is the share of one rater's score that reflects the person being rated rather than that rater's own slant. On the original form it ranged from .09 to .20 by domain, so a domain score needed 10 to 23 evaluators to reach the common .70 benchmark (Table 2 and Figure 2). Removing the three weak items barely changed this, which showed that the issue was not a few bad items but how clearly the items as a whole described observable behavior.

TABLE 2

Stage 1 reliability of the original 28-item form

DOMAIN	SELF ALPHA, ALL ITEMS	SELF ALPHA, REVERSE ITEMS REMOVED	SINGLE-RATER RELIABILITY	RATERS NEEDED FOR .70	RATERS NEEDED FOR .80
People	.73	.80	.14	14	24
Ideas	.70	.78	.09	23	39
Execution	.89	.89 (none removed)	.20	10	17

Note. N = 669 people and 8,340 evaluator ratings. Raters needed are projected from multilevel variance components.

The original form needed 10 to 23 evaluators to reach .70

Reliability of the average evaluator rating, by domain (original 28-item form, 669 people)

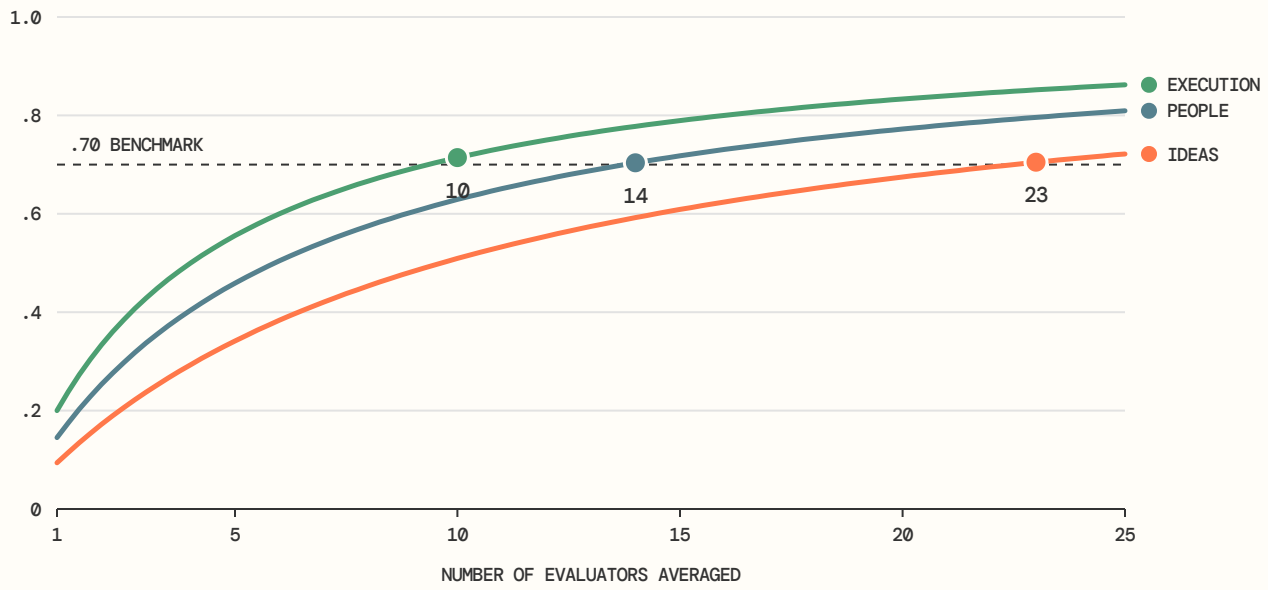


FIGURE 2 **Reliability rises as evaluator ratings are averaged. Dots mark where each domain first reaches .70.**
Stage 1 report (Dr. Harel): variance components from multilevel models, original 28-item form, 669 people, 8,340 ratings

2.4

Ratings bunched near the top

Evaluators were generous, averaging 7.6 to 8.3 out of 10 by domain. A common rule of thumb flags a ceiling effect when more than 15% of respondents give the top score (Terwee et al., 2007); applied to individual items, it flagged 24 of the 25 positively worded items. A crowded top end makes it harder to

tell strong performers apart and leaves little room to show growth on a retake. The scale's labels, which asked how often rather than how well, likely contributed.

2.5

Self-ratings ran lower than evaluator ratings

People rated themselves lower than their evaluators did in every domain: by 0.8 points in Execution, 1.4 in People, and 1.7 in Ideas. In Ideas, at least 90% of people rated themselves below their evaluators' average. The size of the gap did not differ by age, gender, career stage, or whether the person managed others. For users, this gap is one of the most useful things a 360 produces and a natural starting point for a debrief.

2.6

What this meant for the revision

Stage 1 set four rules for the revision: every item positively worded, every item a concrete behavior others can observe, one behavior per item, and a rating scale that asks how well rather than how often. The new scale reserves 10 for "Exceptionally well (one of the best I know)" to give the top more room, since the labels on a scale shape how scores spread out (Bracken et al., 2016). Table 3 summarizes how the revised form differs from the original.

TABLE 3

The original and revised PIE360 at a glance

	ORIGINAL 28-ITEM PIE360	REVISED PIE360
Items	28 (10 People, 9 Ideas, 9 Execution)	36 (12 per domain)
Rating scale	1 to 10, labeled "Occasionally" to "Regularly" (how often)	1 to 10: 1 = Not at all, 4 = Somewhat, 7 = Quite well, 10 = Exceptionally well (one of the best I know) (how well)
Reverse-worded items	3	None
Items that assume the person manages others	Not applicable	None, by design
Forms	Self and evaluator; some pairs asked different things	Self ("I") and rater ("[NAME]") with matched wording
Evidence base	Real-world use (Stage 1)	Stages 1 to 4

SECTION 03

03

Stage 2: Expert review and literature grounding

Stage 2 decided what the revised PIE360 should measure and anchored each part of it in established research.

STAGE 1	STAGE 2	STAGE 3	STAGE 4
Real-world data	Expert review	Statement sorts	National test
P. 10	P. 14	P. 17	P. 21

3.1

The expert panel

The panel had three members, all with doctorates and management experience: Dr. Suzy Welch, who developed the People, Ideas, and Execution framework; Dr. Daniel B. Davis, a sociologist of work and careers; and Dr. Daphna Harel, a statistician. The panel revised the items over several rounds, alongside every later stage.

3.2

Core constructs

A construct is the specific quality an item is meant to capture, such as empathy or timeliness. The panel identified 21 core constructs (8 in People, 5 in Ideas, and 8 in Execution), each measured by one item, and matched each to a counterpart in published research so that every item stands on an established idea. Table 1 in Section 1 lists the research behind all 36 final items: 21 core constructs and the 15 added in the step described next. Some matches are close, where research measures the same behavior or a central part of it. Others are related, where research studies a broader quality or one seen from a different vantage point, such as an inner state that PIE360 rates as a behavior others can see.

3.3

Checking coverage against job-performance research

To make sure nothing important about work was missing, the panel compared the core constructs with seven well-established models of job performance: Campbell (1990), Borman and Motowidlo (1993), Organ (1997), Pulakos, Arad, Donovan, and Plamondon (2000), Bartram's Great Eight competencies (2005), Griffin, Neal, and Parker (2007), and Koopmans et al. (2011). Gaps were filled with constructs that a coworker could observe and that fit a single domain, worded with guidance from research on specific behaviors such as constructive voice (Liang, Farh, & Farh, 2012) and turning ideas into action (Janssen, 2000). Fifteen of the final 36 items measure these added constructs, including helping others, handling disagreement, problem framing, flexible thinking, prioritizing, and quality of work.

A few areas were left out on purpose: skills specific to one job, supervising others (which applies only to managers), behavior aimed at the wider organization (which coworkers rarely see), negative behavior (which would require negatively worded items), and qualities such as persuasion that straddle two domains. Appendix A maps each framework to the final items.

3.4

Design rules

The panel adopted five design rules:

- **Plain, observable wording.** Each item describes a behavior a coworker can see.
- **No reverse-worded items.** They did not fit in Stage 1, and research cautions that they can confuse respondents and blur a scale's structure (Suárez-Álvarez et al., 2018), although some researchers favor careful use over avoidance (Weijters & Baumgartner, 2012).
- **One behavior per item.**
- **No items that assume the person manages others,** so anyone can be rated.
- **Equal item counts across domains,** so no domain's score rests on more questions than another's.

3.5

From constructs to a tested item pool

The panel drafted and reviewed a pool of items, cutting those that overlapped another item or domain, which left 36 statements for the sorting tests in Stage 3. After the sorts, six constructs from a second literature review (recognizing others, courtesy, analytical thinking, reasoned ideas, following procedures, and staying organized) and alternate wordings of three items were added, for 45 items in all to test in Stage 4.

SECTION 04

04

Stage 3: Statement sorts and web probing

A statement can rest on solid research and still be read differently from how its authors intended. Stage 3 tested how people actually read the 36 draft statements before they were fielded.

STAGE 1	STAGE 2	STAGE 3	STAGE 4
Real-world data	Expert review	Statement sorts	National test
P. 10	P. 14	P. 17	P. 21

4.1

Participants and procedure

Participants were 48 undergraduate and graduate students who volunteered to take an anonymous online survey. Nine who matched fewer than half of the statements to the intended definition, a sign they had misunderstood the task, were excluded, leaving 39; most (31) had at least a year of paid work experience. Sorting tests whether wording is read as intended, which takes careful reading rather than knowledge of a particular coworker. Stage 4 then tested the items with working adults.

The survey had three parts:

- **Sort 1, domain fit.** Participants placed each statement, with all three domains shuffled together, in People, Ideas, or Execution, or marked it Unclear.
- **Sort 2, definition match.** Within each domain, participants matched each statement to one of 12 short definitions.
- **Open-ended probes.** Participants explained in their own words how they read key phrases and what a rating of 10 would look like.

This follows a standard method for checking that items fit their intended category before field testing (Anderson & Gerbing, 1991). A statement passed when at least 75% of participants gave the intended answer on both sorts, far above chance (about 33% on Sort 1 and 8% on Sort 2). The open-ended answers were coded by three AI models working independently, which agreed on 90% of answers; Dr. Davis resolved the rest.

4.2

Results

Most statements landed where they were meant to (Table 4), and 24 of the 36 passed both sorts. Most of the rest fell short by drifting toward a neighboring domain, in a consistent pattern. Ideas statements about acting on ideas, such as developing ideas into workable solutions, drifted toward Execution, and an Execution statement about taking responsibility for mistakes drifted toward People.

TABLE 4

Where participants placed statements in Sort 1, by intended domain

INTENDED DOMAIN	PLACED IN PEOPLE	PLACED IN IDEAS	PLACED IN EXECUTION	UNCLEAR	MATCHED INTENDED DEFINITION (SORT 2)
People (12 statements)	86%	6%	8%	1%	89%
Ideas (12 statements)	8%	75%	15%	2%	86%
Execution (12 statements)	12%	3%	83%	2%	86%

Note. N = 39. Each row averages the shares across that domain's 12 statements. Rows may not sum to 100% because of rounding.

The probes explained several of these results and caught problems the sorts could not. In a statement about interacting with others "in an appropriate way," a third of participants read "appropriate" as mere politeness rather than adapting to the setting. "Important responsibilities" had no shared meaning, and "without needing supervision" assumed the person has a boss. Most participants said peers can readily notice behaviors such as revising one's ideas or drawing quieter people into a conversation, which supports rating them from the outside.

4.3

What changed

Statements that fell short, and a few that the probes flagged, were reworded to stay in their intended domain; two that just missed the cutoff were left for Stage 4 to decide. No statement was moved to a different domain. Most rewordings named the intended behavior more directly. For example, "brings a constructive energy to the group" became "brings positive energy to the group," because "constructive" pulled some readers toward ideas and critique, and "handles important responsibilities without needing supervision" became "handles their responsibilities without needing reminders." Table 5 summarizes the 12 reworded statements that appear in the final form.

TABLE 5

Statements reworded after the sorts that appear in the final form

FINAL ID	CONSTRUCT	WHAT THE SORTS OR PROBES SHOWED	WHAT THE REWORDING CHANGED
P5	Interpersonal Adaptability	"Appropriate" often read as merely polite	Named the behavior: adapting one's style to different people
P6	Interpersonal Consistency	Some read it as staying calm under stress	Framed around treating others well on a hard day
P8	Positive Energy*	23% placed it in Ideas	Replaced "constructive" with "positive"
P9	Constructive Disagreement	72% placed it in People	Focused on keeping disagreement about the issue
I2	Discussion Engagement	72% matched the intended definition	Focused on engaging with the ideas under discussion
I4	Discussion Readiness*	21% placed it in Execution	Replaced "prepared," which read as diligence, with thinking the topic through
I7	Problem Framing	23% placed it in Execution	Focused on reframing problems to open up new ideas
I8	Idea Development	31% placed it in Execution	Focused on refining ideas rather than carrying them out
I10	Continuous Learning	31% placed it in Execution	Focused on learning about ideas rather than skills
E3	Self-Management	"Important responsibilities" meant different things; "supervision" assumed a boss	Removed both phrases
E7	Accountability*	56% placed it in People	Focused on owning the results of one's work
E8	Persistence	74% matched the intended definition	Replaced "follows through," shared with another item

Note. * Fielded alongside an alternate wording; Stage 4 chose between them.

SECTION 05

05

Stage 4: Confirmatory factor analysis

Stage 4 tested whether the items group into the three intended domains when working adults rate a real coworker. In a national sample of 405 raters, the final 36 items fit a three-domain model well, every item was strongly tied to its own domain, and the items worked the same way across every rater group tested. Dr. Harel conducted the analysis.

STAGE 1	STAGE 2	STAGE 3	STAGE 4
Real-world data	Expert review	Statement sorts	National test
P. 10	P. 14	P. 17	P. 21

5.1

Sample

Raters were U.S. adults recruited in October 2026 through Prolific, an online research platform, with the sample matched to the U.S. population by sex, age, and ethnicity. Each had worked with a coworker for at least three months and rated the coworker they interacted with most, identified only by initials. After standard quality screens (attention checks and checks for careless or duplicate responses), 405 raters remained (Table 6). Most were peers of the person they rated, and half had worked with that person for more than three years.

Each person in this study was rated by one rater, so Stage 4 tests the structure of the items. Agreement among several raters of the same person will come from real multi-rater use (Section 6).

TABLE 6

Stage 4 sample (N = 405)

CHARACTERISTIC	CATEGORY	N
Rater sex	Female / Male	213 / 192
Rater age	Mean 45.1 (SD 15.1), range 18 to 80	405
Rater ethnicity (Prolific categories)	White / Black / Mixed / Other / Asian	252 / 46 / 44 / 34 / 29
Rater employment (Prolific record)	Full-time / Part-time / Not employed or not in paid work / Other / Not on record	227 / 77 / 41 / 14 / 46
Working relationship	Peer / Rating their manager / Manager rating their report / Other	257 / 73 / 62 / 13
Time worked together	More than 3 years / 1 to 3 years / 7 to 12 months / 3 to 6 months	206 / 118 / 47 / 34
How often they interact	Daily / Several times a week / About weekly / Less often	239 / 141 / 22 / 3
Rated person supervises others	Yes / No / Not sure	166 / 238 / 1

Note. Sex, age, ethnicity, and employment come from Prolific's records of the rater; the other rows are the rater's answers about the person rated.

5.2

Analysis

The fielded pool had 45 items, 15 per domain, including alternate wordings of three items so the data could choose between them. The main analysis was a confirmatory factor analysis, which tests a stated hypothesis about structure: here, that each item reflects one of three related domains. It reports how well that hypothesis reproduces the pattern of correlations among the items (model fit) and how strongly each item reflects its domain (its loading, where 1 would be a perfect reflection and values above .70 are considered strong). Appendix B gives the technical details.

5.3

Choosing the final 36 items

Each item was judged on four things together: how strongly it loaded on its own domain, whether it also leaned toward another domain, whether it overlapped another item, and what content would be lost without it. No item was removed on a single statistic.

Trimming the pool to 11 items per domain gave the best statistical fit (Table 7). The team then restored one item per domain to keep important content: composure in tense moments (P2), originality of ideas (I5), and keeping work organized (E12). Each loaded well, and the small cost in fit was judged worth the broader coverage. Of the nine items cut, three were the weaker of two alternate wordings, four were read by raters as belonging mainly to another domain, and two overlapped closely with items that were kept.

TABLE 7

Model fit across item-selection stages

VERSION	ITEMS	CFI	TLI	RMSEA	SRMR
Full fielded pool	45	.908	.903	.065	.045
Shortest benchmark (11 per domain)	33	.946	.942	.058	.037
Revised PIE360 (12 per domain)	36	.939	.935	.059	.039

Note. N = 405. Higher CFI and TLI and lower RMSEA and SRMR indicate better fit.

Measured against Bartram's (2005) Great Eight competencies, the final form fully covers five and partly covers a sixth. Raters tended to see taking initiative and pursuing ambitious goals as part of how a person engages with ideas, which makes those themes natural candidates for future items.

5.4

Final model results

The final 36-item model fit well: robust CFI = .939, TLI = .935, RMSEA = .059, and SRMR = .039, values that indicate good fit by commonly used standards (Hu & Bentler, 1999). Every item was strongly tied to its own domain, with loadings from .72 to .87 (Figure 3), and each domain was highly reliable, with alpha of .95 to .96 (Table 8).

The three domains were closely related but statistically distinct, with correlations of .87 to .89. Someone rated highly in one domain tends to be rated highly in the others, so results are best read as a profile rather than as three independent traits.

TABLE 8

Reliability and relationships among the domains

DOMAIN	ALPHA	OMEGA-TOTAL	CORRELATION WITH IDEAS [95% CI]	CORRELATION WITH EXECUTION [95% CI]
People	.95	.95	.88 [.84, .92]	.87 [.83, .92]
Ideas	.96	.96		.89 [.85, .92]
Execution	.96	.96		

Note. N = 405. Correlations are between latent factors in the final model.

All 36 items load strongly on their own domain

Standardized loadings in the final model range from .72 to .87

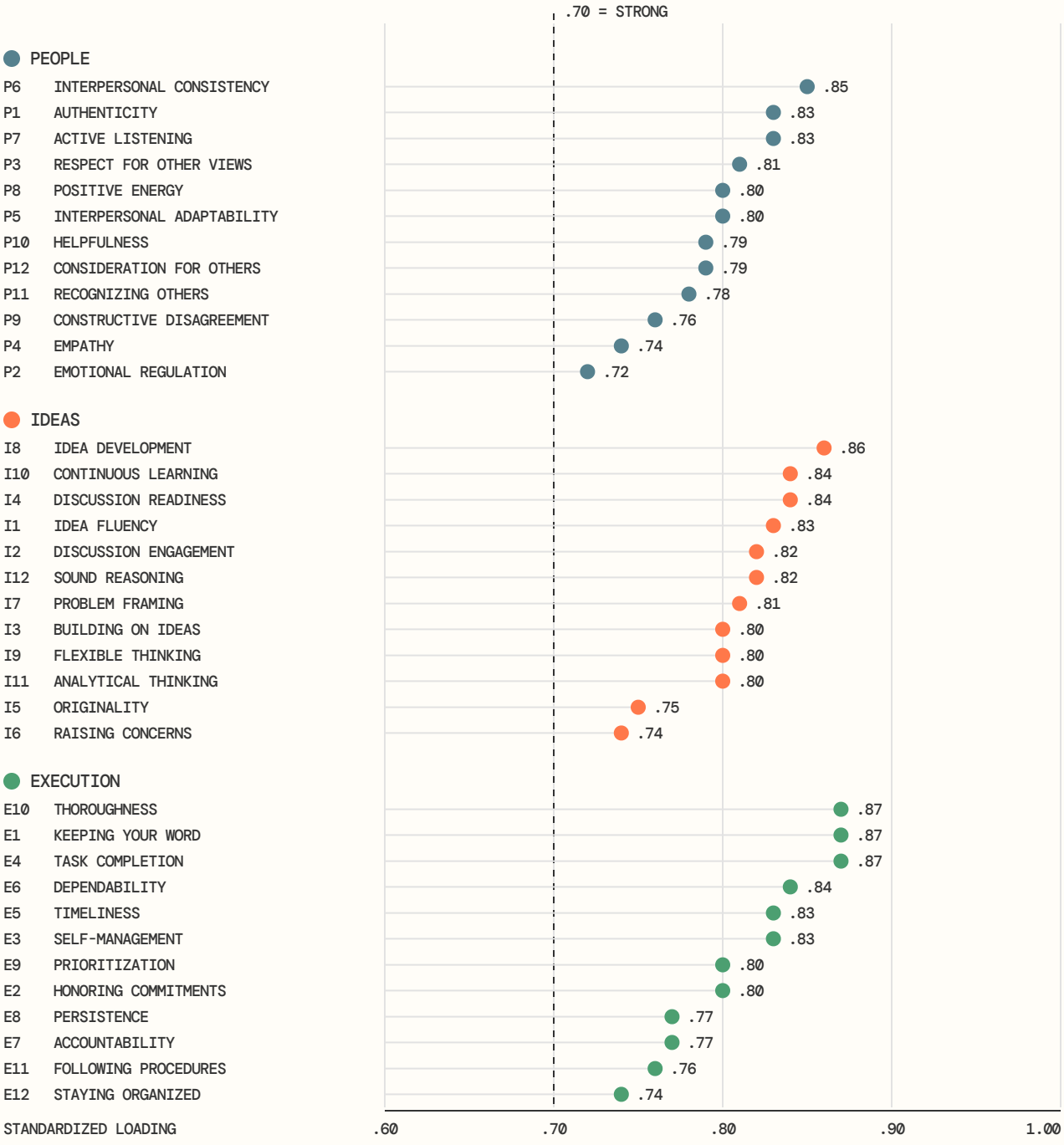


FIGURE 3 **Standardized loadings for the 36 final items, sorted within each domain. Labels are the final construct names.**

Stage 4 final 12-12-12 confirmatory factor analysis, N = 405, final item IDs

5.5

Why the domains overlap

High correlations among domains are common in workplace ratings, because a rater's overall impression of a person colors every rating, an effect known as halo. Two checks confirm that this is what drives the overlap here. A model that separates a general impression from the three domains found that more than 85% of what the items share reflects that overall impression. And careless responding does not explain it: dropping the 75 raters whose answers varied least barely changed the overlap (an average correlation of .86 before and .83 after), and those raters averaged about 8.5 out of 10, so they were most likely describing coworkers they genuinely rated highly. A model that treated ratings as ordered categories reached the same conclusions.

This pattern is well documented. Large studies of job-performance ratings find that a general factor accounts for much of the variance in ratings and that it reflects real differences in overall effectiveness as well as rater tendencies (Scullen, Mount, & Goff, 2000; Viswesvaran, Schmidt, & Ones, 2005). Averaging several raters reduces each rater's personal tendencies, one more reason PIE360 asks for at least 10.

5.6

Fairness checks

A fair instrument measures the same thing in the same way for different groups, so that differences in scores reflect the people rated rather than how the items work. This property, called measurement invariance, was tested by holding the model's measurement identical across groups and checking how much fit suffered. In every comparison the loss was negligible by the usual standard (Cheung & Rensvold, 2002), including rater sex, rater ethnicity, the working relationship, and how long and how often rater and coworker worked together (Table 9). Smaller groups, such as specific ethnic groups, will be examined as more data come in.

Older raters gave slightly higher Execution ratings (standardized effect = .12), with no difference on People or Ideas, and no individual item worked differently by age.

TABLE 9

Fairness checks at a glance

COMPARISON	GROUPS (N)	CHANGE IN CFI WHEN MEASUREMENT IS HELD EQUAL	VERDICT
Rater sex	Women 213, men 192	-.001	Supported
Rated person supervises others	Yes 166, no 238	-.001	Supported
How often they interact	Daily 239, less often 166	+.001	Supported
Time worked together	1 year or less 81, more than 1 year 324	+.001	Supported
Working relationship	Peers 257, manager-report pairs 135	-.001	Supported
Rater ethnicity	White 252, non-White 153	-.002	Supported (broad comparison)

Note. Change in CFI when loadings and item intercepts are held equal across groups; a change smaller than .01 is considered negligible.

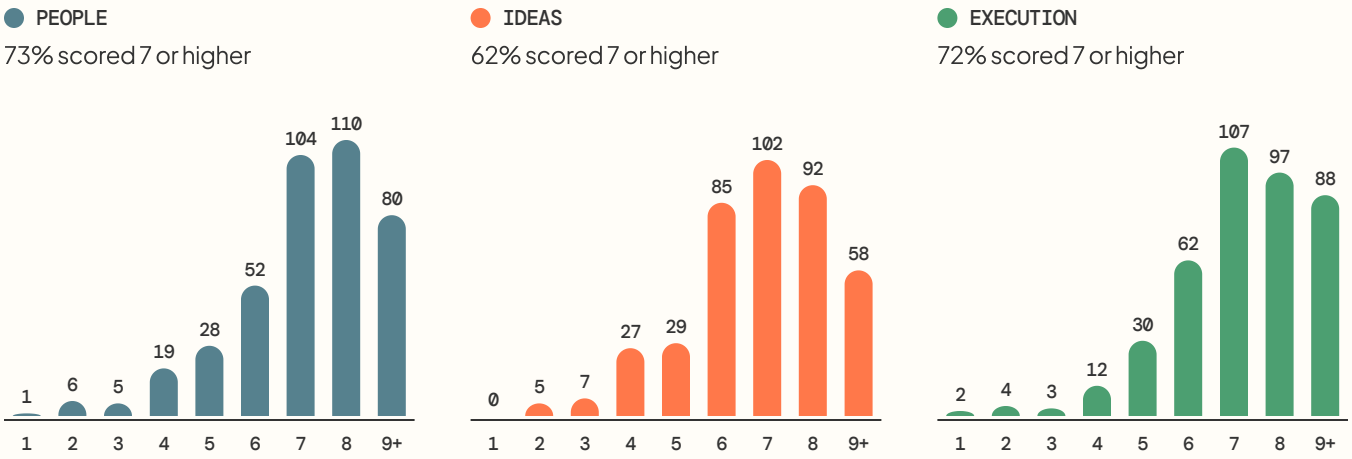
5.7

How the new scale behaved

The relabeled scale gave the top end more room. In Stage 4, the share of raters choosing 10 on an item ranged from 8% to 25% (median 15%), and half of the items stayed within the 15% ceiling guideline, compared with 1 of 25 on the original form. Fewer than 4% of people received a perfect domain score. Ratings still leaned toward the upper half of the scale (Figure 4); ratings from real users of the revised form will show how the scale performs in everyday use.

Most domain scores fall between 7 and 10

Number of raters by average domain rating (1 to 10 scale), final 36 items, N = 405



Bars group average ratings into one-point bands: 1 means 1.0 to 1.9, and 9+ means 9.0 to 10.

FIGURE 4 **How domain scores spread across the scale in Stage 4. Ratings lean toward the upper half of the scale while still spreading across it.**

Stage 4 data, clean N = 405; domain score = average of the rater's 12 item ratings in that domain

SECTION 06

06

What the results mean and what comes next

Together, the four stages show that the revised PIE360 does what it was designed to do. Its 36 items describe observable behaviors grounded in research, people read them as intended, and in a national sample they formed three reliable domains that held up across rater groups. Here is what that means in practice.

6.1	6.2	6.3	6.4	6.5
Read the three domains as a profile	Invite at least 10 raters	Pooling all raters is supported	Use PIE360 for development	Next steps
P. 30	P. 30	P. 31	P. 31	P. 31

6.1

Read the three domains as a profile

People, Ideas, and Execution are distinct but closely related. Someone rated highly in one domain is usually rated highly in the others, partly because raters form an overall impression of a person and partly because effective people tend to be effective in more than one way. The most useful way to read a PIE360 report is as a profile: which domain stands out for this person relative to the others, and which specific behaviors raters highlighted, rather than three scores read in isolation.

6.2

Invite at least 10 raters

Averaging many raters is what makes a 360 trustworthy, because each rater's personal slant shrinks as more views are added. PIE360 releases results only after at least 10 raters respond, and published research supports that minimum. Table 10 uses published estimates of single-rater reliability in workplace ratings (Conway & Huffcutt, 1997) to show how reliability rises as raters are added. With 10 raters, it reaches .81 to .91 depending on who is rating, above the common .70 benchmark and in most cases above .80. An often-cited recommendation of about 4 supervisors, 8 peers, or 9 direct reports (Greguras & Robie, 1998, as summarized by Bracken et al., 2001) also falls at or below 10.

TABLE 10

How reliability grows as raters are added

NUMBER OF RATERS	SUPERVISORS	PEERS	DIRECT REPORTS
1	.50	.37	.30
5	.83	.75	.68
10	.91	.85	.81
15	.94	.90	.87

Note. Single-rater reliabilities from Conway and Huffcutt (1997); values for more raters are projected with the Spearman-Brown formula. A reliability of .70 is generally considered acceptable and .80 good.

The original form needed more raters than these published figures suggest, one reason the revised items were written to make each rating more informative. Users already invite plenty of raters: on the original form, the median user invited 12. Inviting a mix of people who know the person's work well gives the fullest picture.

6.3

Pooling all raters is supported

PIE360 combines all raters into one report rather than breaking results out by relationship, which protects raters' anonymity and encourages candor. Stage 4 supports that design: the items worked the same way for peers and for manager-report pairs, and for raters who knew the person for shorter or longer periods.

6.4

Use PIE360 for development

PIE360 is built for growth and self-understanding. Research on 360 feedback shows that results lead to the most improvement when people set specific goals and talk them through, ideally with a coach (Smither et al., 2003, 2005). PIE360 reports are designed for that kind of conversation, and the gap between how people see themselves and how others see them is often the best place to start: in Stage 1, most users rated themselves lower than their raters did. PIE360 is intended for development, not for decisions about pay, promotion, or hiring.

6.5

Next steps

This paper reports the first validation of the revised PIE360. The research continues in three ways:

1. **Ongoing checks with real users.** As people use the revised form, we will keep checking how it performs: how closely raters agree, how many raters a stable score needs, how self-ratings compare with others' ratings on the new items, and how ratings from the clients, mentors, friends, and family members that users invite compare with ratings from coworkers.
2. **Links to outcomes.** Future studies will relate PIE360 scores to other measures of work performance and to growth over time, including change between retakes.
3. **The full report.** The report's summary features, including the overall score, the self-awareness score, and the archetypes, build on the three domain scores. Future work will examine these features directly and develop norms for reading them.

Each step will be reported in a future research report.

References

Anderson, J. C., & Gerbing, D. W. (1991). Predicting the performance of measures in a confirmatory factor analysis with a pretest assessment of their substantive validities. *Journal of Applied Psychology, 76*(5), 732–740. https://doi.org/10.1037/0021-9010.76.5.732

Atwater, L. E., Ostroff, C., Yammarino, F. J., & Fleenor, J. W. (1998). Self-other agreement: Does it really matter? *Personnel Psychology, 51*(3), 577–598. https://doi.org/10.1111/j.1744-6570.1998.tb00252.x

Bartram, D. (2005). The Great Eight competencies: A criterion-centric approach to validation. *Journal of Applied Psychology, 90*(6), 1185–1203. https://doi.org/10.1037/0021-9010.90.6.1185

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bodie, G. D. (2011). The Active-Empathic Listening Scale (AELS): Conceptualization and evidence of validity within the interpersonal domain. *Communication Quarterly, 59*(3), 277–295. https://doi.org/10.1080/01463373.2011.583495

Borman, W. C., & Motowidlo, S. J. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 71–98). Jossey-Bass.

Bracken, D. W., & Rose, D. S. (2011). When does 360-degree feedback create behavior change? And how would we know it when it does? *Journal of Business and Psychology, 26*(2), 183–192. https://doi.org/10.1007/s10869-011-9218-5

Bracken, D. W., Rose, D. S., & Church, A. H. (2016). The evolution and devolution of 360° feedback. *Industrial and Organizational Psychology, 9*(4), 761–794. https://doi.org/10.1017/iop.2016.93

Bracken, D. W., Timmreck, C. W., Fleenor, J. W., & Summers, L. (2001). 360 feedback from another angle. *Human Resource Management, 40*(1), 3–20. https://doi.org/10.1002/hrm.4012

Campbell, J. P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 1, pp. 687–732). Consulting Psychologists Press.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Claessens, B. J. C., van Eerde, W., Rutte, C. G., & Roe, R. A. (2007). A review of the time management literature. *Personnel Review, 36*(2), 255–276. https://doi.org/10.1108/00483480710726136

Conway, J. M., & Huffcutt, A. I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance, 10*(4), 331–360. https://doi.org/10.1207/s15327043hup1004_2

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Psychological Assessment Resources.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology, 44*(1), 113–126. https://doi.org/10.1037/0022-3514.44.1.113

De Dreu, C. K. W., & Weingart, L. R. (2003). Task versus relationship conflict, team performance, and team member satisfaction: A meta-analysis. *Journal of Applied Psychology, 88*(4), 741–749. https://doi.org/10.1037/0021-9010.88.4.741

De Jong, J., & Den Hartog, D. (2010). Measuring innovative work behaviour. *Creativity and Innovation Management, 19*(1), 23–36. https://doi.org/10.1111/j.1467-8691.2010.00547.x

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology, 92*(6), 1087–1101. https://doi.org/10.1037/0022-3514.92.6.1087

Ferris, G. R., Treadway, D. C., Kolodinsky, R. W., Hochwarter, W. A., Kacmar, C. J., Douglas, C., & Frink, D. D. (2005). Development and validation of the Political Skill Inventory. *Journal of Management, 31*(1), 126–152. https://doi.org/10.1177/0149206304271386

Fleenor, J. W., Smither, J. W., Atwater, L. E., Braddy, P. W., & Sturm, R. E. (2010). Self–other rating agreement in leadership: A review. *The Leadership Quarterly, 21*(6), 1005–1034. https://doi.org/10.1016/j.leaqua.2010.10.006

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly, 73*(2), 349–360. https://doi.org/10.1093/poq/nfp031

Greguras, G. J., & Robie, C. (1998). A new look at within-source interrater reliability of 360-degree feedback ratings. *Journal of Applied Psychology, 83*(6), 960–968. https://doi.org/10.1037/0021-9010.83.6.960

Griffin, M. A., Neal, A., & Parker, S. K. (2007). A new model of work role performance: Positive behavior in uncertain and interdependent contexts. *Academy of Management Journal, 50*(2), 327–347. https://doi.org/10.5465/amj.2007.24634438

Gross, J. J. (1998). The emerging field of emotion regulation: An integrative review. *Review of General Psychology, 2*(3), 271–299. https://doi.org/10.1037/1089-2680.2.3.271

Guilford, J. P. (1967). *The nature of human intelligence*. McGraw-Hill.

Hargadon, A. B., & Bechky, B. A. (2006). When collections of creatives become creative collectives: A field study of problem solving at work. *Organization Science, 17*(4), 484–500. https://doi.org/10.1287/orsc.1060.0200

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Janssen, O. (2000). Job demands, perceptions of effort-reward fairness and innovative work behaviour. *Journal of Occupational and Organizational Psychology, 73*(3), 287–302. https://doi.org/10.1348/096317900167038

Jehn, K. A. (1995). A multimethod examination of the benefits and detriments of intragroup conflict. *Administrative Science Quarterly, 40*(2), 256–282. https://doi.org/10.2307/2393638

Kahn, W. A. (1990). Psychological conditions of personal engagement and disengagement at work. *Academy of Management Journal, 33*(4), 692–724. https://doi.org/10.5465/256287

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Koopmans, L., Bernaards, C. M., Hildebrandt, V. H., Schaufeli, W. B., de Vet, H. C. W., & van der Beek, A. J. (2011). Conceptual frameworks of individual work performance: A systematic review. *Journal of Occupational and Environmental Medicine, 53*(8), 856–866. https://doi.org/10.1097/JOM.0b013e318226a763

Koopmans, L., Bernaards, C., Hildebrandt, V., van Buuren, S., van der Beek, A. J., & de Vet, H. C. W. (2013). Development of an individual work performance questionnaire. *International Journal of Productivity and Performance Management, 62*(1), 6–28. https://doi.org/10.1108/17410401311285273

Liang, J., Farh, C. I. C., & Farh, J.-L. (2012). Psychological antecedents of promotive and prohibitive voice: A two-wave examination. *Academy of Management Journal, 55*(1), 71–92. https://doi.org/10.5465/amj.2010.0176

Martin, M. M., & Rubin, R. B. (1995). A new measure of cognitive flexibility. *Psychological Reports, 76*(2), 623–626. https://doi.org/10.2466/pr0.1995.76.2.623

National Center for O*NET Development. (n.d.). *Analyzing data or information* (Work activity 4.A.2.a.4). O*NET OnLine. Retrieved October 30, 2026, from https://www.onetonline.org/find/descriptor/result/4.A.2.a.4

Nowack, K. M., & Mashihi, S. (2012). Evidence-based answers to 15 questions about leveraging 360-degree feedback. *Consulting Psychology Journal: Practice and Research, 64*(3), 157–182. https://doi.org/10.1037/a0030011

Organ, D. W. (1988). *Organizational citizenship behavior: The good soldier syndrome*. Lexington Books.

Organ, D. W. (1997). Organizational citizenship behavior: It's construct clean-up time. *Human Performance, 10*(2), 85–97. https://doi.org/10.1207/s15327043hup1002_2

Owens, B. P., Baker, W. E., Sumpter, D. M., & Cameron, K. S. (2016). Relational energy at work: Implications for job engagement and job performance. *Journal of Applied Psychology, 101*(1), 35–49. https://doi.org/10.1037/apl0000032

Parker, S. K., & Axtell, C. M. (2001). Seeing another viewpoint: Antecedents and outcomes of employee perspective taking. *Academy of Management Journal, 44*(6), 1085–1100. https://doi.org/10.5465/3069390

Podsakoff, P. M., MacKenzie, S. B., Paine, J. B., & Bachrach, D. G. (2000). Organizational citizenship behaviors: A critical review of the theoretical and empirical literature and suggestions for future research. *Journal of Management, 26*(3), 513–563. https://doi.org/10.1177/014920630002600307

Pulakos, E. D., Arad, S., Donovan, M. A., & Plamondon, K. E. (2000). Adaptability in the workplace: Development of a taxonomy of adaptive performance. *Journal of Applied Psychology, 85*(4), 612–624. https://doi.org/10.1037/0021-9010.85.4.612

Reiter-Palmon, R., Mumford, M. D., O'Connor Boes, J., & Runco, M. A. (1997). Problem construction and creativity: The role of ability, cue consistency, and active processing. *Creativity Research Journal, 10*(1), 9–23. https://doi.org/10.1207/s15326934crj1001_2

Rich, B. L., LePine, J. A., & Crawford, E. R. (2010). Job engagement: Antecedents and effects on job performance. *Academy of Management Journal, 53*(3), 617–635. https://doi.org/10.5465/amj.2010.51468988

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology, 85*(6), 956–970. https://doi.org/10.1037/0021-9010.85.6.956

Simons, T. (2002). Behavioral integrity: The perceived alignment between managers' words and deeds as a research focus. *Organization Science, 13*(1), 18–35. https://doi.org/10.1287/orsc.13.1.18.543

Smither, J. W., London, M., Flautt, R., Vargas, Y., & Kucine, I. (2003). Can working with an executive coach improve multisource feedback ratings over time? A quasi-experimental field study. *Personnel Psychology, 56*(1), 23–44. https://doi.org/10.1111/j.1744-6570.2003.tb00142.x

Smither, J. W., London, M., & Reilly, R. R. (2005). Does performance improve following multisource feedback? A theoretical model, meta-analysis, and review of empirical findings. *Personnel Psychology, 58*(1), 33–66. https://doi.org/10.1111/j.1744-6570.2005.514_1.x

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology, 113*(1), 117–143. https://doi.org/10.1037/pspp0000096

Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology, 89*(2), 342–357. https://doi.org/10.1037/0022-0663.89.2.342

Suárez-Álvarez, J., Pedrosa, I., Lozano, L. M., García-Cueto, E., Cuesta, M., & Muñiz, J. (2018). Using reversed items in Likert scales: A questionable practice. *Psicothema, 30*(2), 149–158. https://doi.org/10.7334/psicothema2018.33

Terwee, C. B., Bot, S. D. M., de Boer, M. R., van der Windt, D. A. W. M., Knol, D. L., Dekker, J., Bouter, L. M., & de Vet, H. C. W. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology, 60*(1), 34–42. https://doi.org/10.1016/j.jclinepi.2006.03.012

VandeWalle, D. (1997). Development and validation of a work domain goal orientation instrument. *Educational and Psychological Measurement, 57*(6), 995–1015. https://doi.org/10.1177/0013164497057006009

Van Dyne, L., & LePine, J. A. (1998). Helping and voice extra-role behaviors: Evidence of construct and predictive validity. *Academy of Management Journal, 41*(1), 108–119. https://doi.org/10.5465/256902

Van Scotter, J. R., & Motowidlo, S. J. (1996). Interpersonal facilitation and job dedication as separate facets of contextual performance. *Journal of Applied Psychology, 81*(5), 525–531. https://doi.org/10.1037/0021-9010.81.5.525

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology, 90*(1), 108–131. https://doi.org/10.1037/0021-9010.90.1.108

Weijters, B., & Baumgartner, H. (2012). Misresponse to reversed and negated items in surveys: A review. *Journal of Marketing Research, 49*(5), 737–747. https://doi.org/10.1509/jmr.11.0368

Zhou, J., & George, J. M. (2001). When job dissatisfaction leads to creativity: Encouraging the expression of voice. *Academy of Management Journal, 44*(4), 682–696. https://doi.org/10.5465/3069410

APPENDIX A

Literature crosswalk

Table A1 maps the dimensions of the seven job-performance frameworks used in Stage 2 to the final 36 items. "Covered" means at least one final item measures the dimension directly; "partly covered" means an item measures part of it. Some dimensions were excluded by design, and two are candidates for future items (Section 5.3).

TABLE A1

Job-performance frameworks and the final PIE360 items

FRAMEWORK	DIMENSION	FINAL ITEMS	COVERAGE
Campbell (1990)	Non-job-specific task proficiency	E4, E5, E10	Covered
Campbell (1990)	Demonstrating effort	E8	Covered
Campbell (1990)	Maintaining personal discipline	E1, E2, E11	Covered
Campbell (1990)	Facilitating peer and team performance	P8, P10	Covered
Campbell (1990)	Management and administration (individual level)	E9, E12	Covered
Campbell (1990)	Job-specific task proficiency	None	Excluded by design: differs by job
Campbell (1990)	Written and oral communication	I2 (in part)	Excluded as a separate skill: spans People and Ideas
Campbell (1990)	Supervision and leadership	None	Excluded by design: applies only to managers
Borman & Motowidlo (1993)	Task and contextual performance	P1 to P12; E1 to E12	Covered
Organ (1997)	Organizational citizenship behavior	P4, P10, P11, P12; E6	Covered
Pulakos et al. (2000)	Handling work stress	P2	Covered
Pulakos et al. (2000)	Interpersonal adaptability	P3, P5	Covered
Pulakos et al. (2000)	Solving problems creatively	I5, I8	Covered
Pulakos et al. (2000)	Learning work tasks and technologies	I10	Covered

FRAMEWORK	DIMENSION	FINAL ITEMS	COVERAGE
Pulakos et al. (2000)	Dealing with uncertainty	I9	Partly covered
Bartram (2005)	Supporting and Cooperating	P1, P3, P4, P6, P10, P11	Covered
Bartram (2005)	Analyzing and Interpreting	I7, I11, I12	Covered
Bartram (2005)	Creating and Conceptualizing	I1, I5, I9, I10	Covered
Bartram (2005)	Organizing and Executing	E4, E5, E9, E10, E12	Covered
Bartram (2005)	Adapting and Coping	P2, I9	Covered
Bartram (2005)	Interacting and Presenting	P5, P9	Partly covered: persuasion and influence excluded because they span People and Ideas
Bartram (2005)	Leading and Deciding	None	Leadership excluded by design; taking initiative a candidate for future items
Bartram (2005)	Enterprising and Performing	None	Candidate for future items
Griffin, Neal, & Parker (2007)	Individual task proficiency	E1 to E10	Covered
Griffin, Neal, & Parker (2007)	Team member proficiency	P7, P10	Covered
Griffin, Neal, & Parker (2007)	Individual task adaptivity	I9	Partly covered
Griffin, Neal, & Parker (2007)	Team member proactivity	I6	Partly covered
Griffin, Neal, & Parker (2007)	Individual task proactivity	None	Candidate for future items
Griffin, Neal, & Parker (2007)	Team member adaptivity	None	Excluded by design: behavioral adaptability spans all three domains
Griffin, Neal, & Parker (2007)	Organization member behaviors	None	Excluded by design: hard for coworkers to observe
Koopmans et al. (2011)	Task performance	E1 to E12	Covered
Koopmans et al. (2011)	Contextual performance	P1 to P12; E6, E8	Covered
Koopmans et al. (2011)	Adaptive performance	P2, I9, I10	Partly covered

FRAMEWORK	DIMENSION	FINAL ITEMS	COVERAGE
Koopmans et al. (2011)	Counterproductive work behavior	None	Excluded by design: all items are positively worded

Note. Final item IDs; construct names are in Table 1. Mapping from the Stage 2 literature review, updated for the items retained in Stage 4.

APPENDIX B

Technical details

Estimation. All Stage 4 models specified three correlated factors, with each item loading on one prespecified domain, and were estimated in R with the lavaan package (Rosseel, 2012), using robust maximum likelihood and full-information handling of missing responses. The ten-point items were treated as approximately continuous; an ordinal (WLSMV) model was run as a check. In the 45-item model, residual correlations between the three alternate-wording pairs were specified in advance. Fit was judged with the robust comparative fit index (CFI), Tucker–Lewis index (TLI), root mean square error of approximation (RMSEA; final model 90% confidence interval .054 to .064), and standardized root mean square residual (SRMR), with commonly cited reference values (CFI and TLI near .95, RMSEA near .06, SRMR below .08; Hu & Bentler, 1999) treated as guides rather than pass–fail cutoffs. Modification indices were used to locate redundant wording and overlap between domains; none was freed automatically. Reliability was estimated with Cronbach's alpha and McDonald's omega-total, which agreed to two decimals. Results were unchanged when 13 raters who missed exactly one attention check but passed every other screen were retained (N = 418).

Fairness analyses. Measurement invariance was tested at the configural, metric, and scalar levels, with a change in CFI no worse than −.01 (Cheung & Rensvold, 2002) and small changes in RMSEA and SRMR (Chen, 2007) as the main criteria. Age was modeled continuously with a multiple-indicator multiple-cause model (Execution effect $p = .009$; People and Ideas not significant), and item-level age effects were screened with Benjamini–Hochberg false-discovery-rate correction (Benjamini & Hochberg, 1995; smallest adjusted $p = .72$).

Analysis record. Stage 1 analyses were conducted by Dr. Harel in R, using multilevel models and the psych package. Stage 3 sort results and probe coding were computed from the survey export. Stage 4 confirmatory factor analyses, reliability, measurement invariance, and age analyses were conducted by Dr. Harel in R with lavaan, and the robustness checks reported in Section 5.5 (ordinal estimation, bifactor model, and careless-responding screens) were reproduced independently from the cleaned data (R 4.3.3, lavaan 0.6.17). Analysis scripts and de-identified data are held by Becoming You Labs and are available to qualified researchers on request.

Research on what people value, and how fully they live it.

Becoming You Labs

RESEARCH@BECOMINGYOULABS.COM ↗