Tilt 365®

EVALUATION OF THE POSITIVE INFLUENCE PREDICTOR

TABLE OF CONTENTS

EXECUTIVE SUMMARY

This is the first psychometric evaluation of the revised version of the *T*ilt 365 Positive Influence Predictor, which was moved to a new platform in February 2018. The Positive Influence Predictor is a multi-rater assessment based on the *T*ilt Framework, which is a wholistic model of character strengths. The purpose of the assessment is to show people how they and others perceive their behavior so that they can learn to balance all 12 core character strengths. We evaluated the evidence of structural validity, internal consistency reliability, inter-rater reliability, and inter-rater agreement.

Using a subset of independent other-ratings, we found evidence for validity based on internal structure. Overall, factor analyses indicated that the Positive Influence Predictor seems to fit the hierarchical model implied by the Tilt Framework. However, the full model should be re-evaluated when an adequate amount of data is available to make parameter estimates trustworthy. Although the results from the full hierarchical factor model should be interpreted with caution, the absolute fit indices and high factor loadings provided evidence for internal structure validity that is similar to previous studies on the prior platform. Paring the full model down into 4 quadrant-level models allowed for reliable estimation with the sample size available. These models had acceptable fit using comparative and absolute fit indices, and all items had high factor loadings onto their respective character strengths.

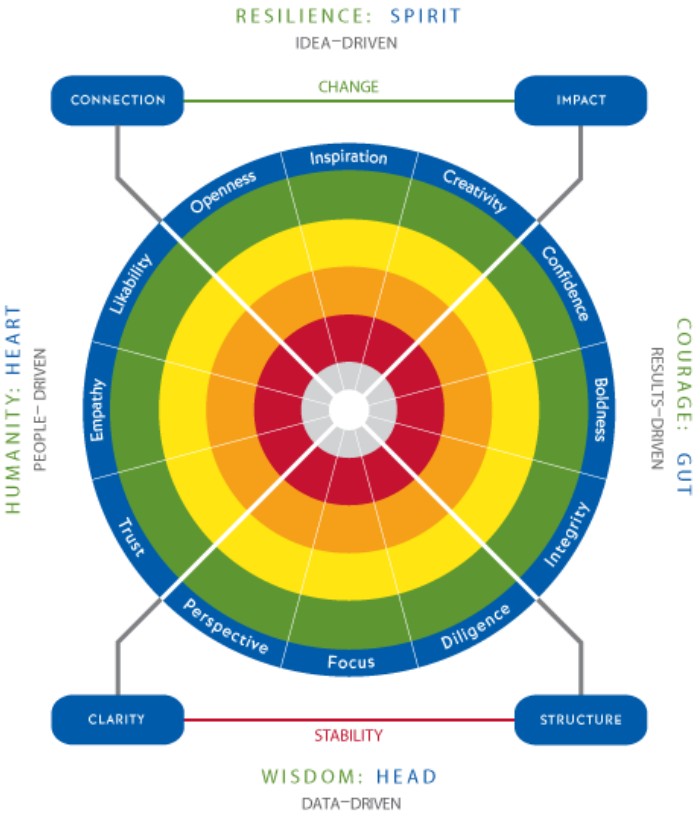
We evaluated the internal consistency reliability of the 12 scales using two different statistics: alpha and omega. Although alpha is the most commonly reported statistic, psychometricians recommend indices based in structural equation modeling like omega. Scale omegas ranged from .65 to .83, and omegas for the quadrants were as high as .91. Only the focus scale was below the typical cutoff of .70. Analyses did not indicate any "bad" items in the scale, so it will be examined further when more data is available.

We found evidence for inter-rater reliability and agreement. All *ICC*(1) values were greater than .05, and most were greater than .10. Estimates of the reliability of mean ratings, using *ICC*(2) values, were all higher than .40. Inter-rater agreement was assessed using r_{WG} for items and $r_{WG(J)}$ for scales. All estimates using the rectangular distribution as the null were greater than .70, and most estimates using the triangular distribution as the null were greater than .70. Average deviation values were all well below the upper limit of 1.5, which indicates high agreement. Collectively this provides evidence for inter-rater reliability and inter-rater agreement.

Overall, the Positive Influence Predictor on the new platform demonstrated acceptable psychometric properties for use as a developmental assessment.

OVERVIEW OF THE *T*ILT FRAMEWORK

The *T*ilt Model draws on research in character science to provide a wholistic framework of 12 core character strengths. The *T*ilt Model is comprised of four quadrants: Resilience, Courage, Wisdom, and Humanity. Each of those quadrants is made up of three character strengths. The purpose of the model is to describe both people's natural strengths and the degree to which they can balance all 12 strengths in a given situation. Natural preference and balance are ascertained with two different assessments: The *T*rue Tilt Personality Profile (TTP) and Positive Influence Predictor, respectively. This report focuses on the assessment of balance using the Positive Influence Predictor. Balance does not necessarily reflect the individual's natural preference; people can (and should) develop all 12 character strengths so that they can use the strengths which are most appropriate for any situation.



Balance

Everyone has a set of character strengths that come more easily to them than others, and if they do not know when to use each strength, then preferred character strengths can be overused and turn into weaknesses. For example, when someone is too "self-assured" it becomes arrogant. Conversely, strengths that are not one of the natural preferences can be underused. For example, not being self-assured enough comes across as insecure. Both overusing and underusing character strengths are detrimental. With self-awareness and conscious effort people can actively develop all 12 of the strengths so that they are able to express each in balance. The purpose of the Tilt Framework is to help people become self-aware so that they can learn to balance all 12 Character Strengths.

Positive Influence Predictor

In order to learn to balance, individuals need an assessment of their current strengths and to monitor how that shifts across time and contexts. The Positive Influence Predictor is a multi-rater assessment that allows individuals to receive feedback from observers who have varying relationships with the them (e.g. boss, peer, direct report). Individuals often lack accurate self-awareness, so external observations can be used to clarify blind spots and misconceptions. In addition, people act differently in different situations and with different people, so requesting observations from varying types of raters (e.g., boss v. direct report) across a variety of contexts can illustrate the different ways people are perceived.

The Positive Influence Predictor uses a bipolar scale to assess how people express character traits. Both the overuse and underuse of commendable traits can be detrimental if too frequent or extreme. Many other assessments only measure the underuse of either a trait or competency. For example, the assessment asked respondents to indicate how "authentic" a person was on a scale of 1 to 7 and 7 represents the ideal amount. The *T*ilt model recognizes that overuse of favorable traits can also lead to harmful behaviors. Thus, the Positive Influence Predictor uses response options that allow observers to indicate whether the trait is used too little or too much. For example, respondents would use the scale below to rate the frequency with whether the ratee was brave.

This ideal is based in the philosophy of Aristotle's golden mean, or the desirable middle ground between two extremes. Thus, the Positive Influence Predictor shows whether people act in a way that demonstrates the appropriate amount of a commendable trait.

The Positive Influence Predictor has 48 of these bipolar rating items and three free response items. Each of the nine response options on the scale is labeled with a statement describing the frequency of expressing the trait. The middle response option indicates balance. It takes about 15 minutes to complete. The Positive Influence Predictor was revised slightly based on the results of the last analysis and moved to a different platform. This is the first evaluation of the revised assessment.

SAMPLE DATA

The sample was comprised of responses to the Positive Influence Predictor from February 2018 to mid-July 2019. The revised version of the assessment was launched in February 2018, so no responses before that time were included. There was a total of 10,467 responses and 913 unique people being rated. Of those, 1,087 responses were removed for careless responding. Careless responding was assessed using three post hoc indicators: Psychometric Synonyms, Even-Odd Consistency, and Mahalanobis distance (for more information on detecting careless responding see Curran, 2016; Meade & Craig, 2012). These are continuous indicators, so, consistent with previous research, a cut score of two standard deviations below the mean (Psychometric Synonyms & Even-Odd Consistency) or

two standard deviations above the mean (Mahalanobis Distance) was developed for each indicator (Francavilla, Meade, & Young, 2018).

Some analyses in this report are predicated on the assumption of independent observations (e.g., factor analysis). Because multi-rater data is dependent by its nature (i.e., the ratings of the same target individual are dependent), we randomly selected a subset of 888 responses with no missing values out of the 9,380 responses. These contained only 1 rating per target individual, none of which were self-ratings. Self-ratings were excluded because they are categorically different than other-ratings and are not used in balance calculations. Approximately 37% of the randomly selected ratings were peers, 24% were direct reports, 15% were managers, 11% were friends, 5% were clients, and 9% were in the Other category.

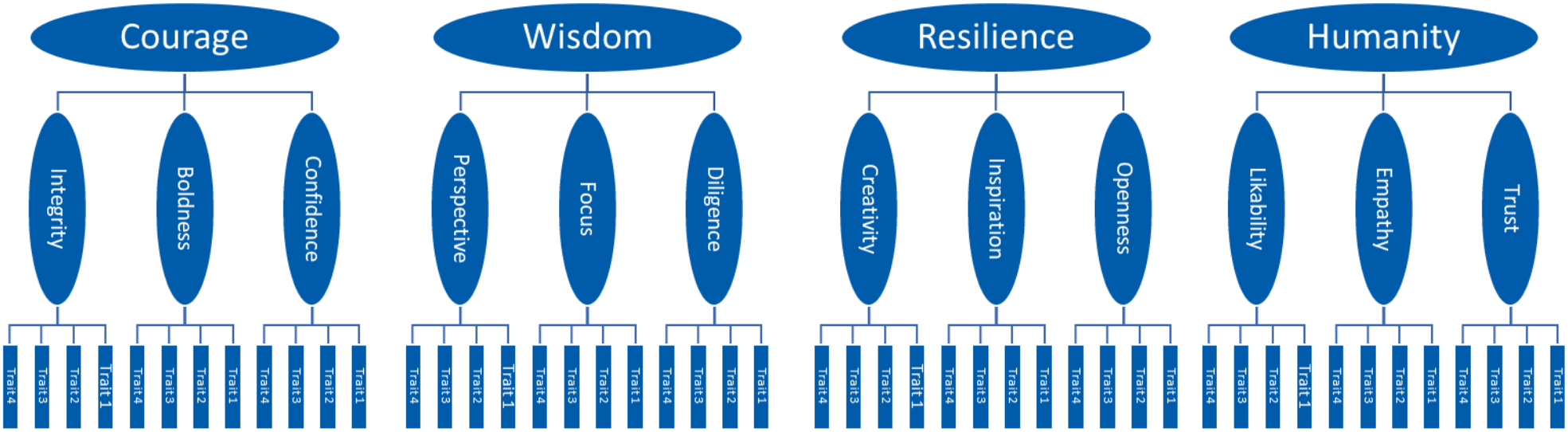
INTERNAL STRUCTURE VALIDITY EVIDENCE

Validity refers to the degree that evidence supports an assessment's appropriateness for predicting or drawing inferences about certain criteria. Because the emphasis is on relationship with another construct, validity is not an inherent property of an assessment. This means that a test can be valid for one purpose, but not for another. In addition, validity is not all or nothing; it is a continuum. Thus, an assessment can never be called a "valid assessment," but one can only provide evidence that it is valid enough for a given purpose. There are different ways to provide validity evidence, and the appropriate type to assess depends on the intended use of the assessment. According to *Standards for Educational and Psychological Testing* (2014) there are four types of validity evidence. We assessed evidence based on internal structure. Validity evidence based on internal structure is the evaluation of the degree the relationships between assessment items and components conform to the conceptual framework of the construct(s) being assessed. The Positive Influence Predictor is based on the *T*ilt Framework, so the structure of the assessment should match the implied structure of the *T*ilt Framework in order to show evidence for validity based on internal structure.

Factor Structure

The structure implied by the *T*ilt Model and measured by the Positive Influence Predictor is a second-order factor model. The first-order factors are the 12 character strengths, which are each measured by responses to four items. The second-order factors are the four quadrants, which are each measured by three first-order factors (i.e., character strengths). For example, Humanity is a second-order factor comprise of the character strengths likability, empathy, and trust, and each of those strengths are measured with four items representing character traits. The structure is depicted on the figure on the next page.

Figure 1. **Theoretical Factor Structure of the *T*ilt Model**



Confirmatory Factor Analysis (CFA)

We evaluated five CFA models to assess internal structure validity evidence. First, we fit the hierarchical model implied by the *T*ilt framework. However, we only had a sample size of 888 unique ratings, and the hierarchical model has over 100 estimated parameters. The recommended sample-size-to-parameters ratio is 1:20, and as the ratio falls below 10:1 the trustworthiness of the results decreases (Kline, 2016). The ratio for the full hierarchical model was less than 1:8. The hierarchical model was able to converge, so it is reported, but we supplemented with the results of a separate CFA for each quadrant. These CFAs each estimated 12 error variances, 12 loadings, and three covariances between the three latent strengths, totaling 27 estimated parameters. With an approximately 1:33 sample-size-to-parameters ratio, the estimates from these models are more trustworthy than those of the full model.

Model Fit

There are several ways to evaluate whether or not the proposed model adequately describes the data, but there is no way to prove with absolute certainty that a proposed model fits the data. Different indices of model fit provide different types of evidence either for or against a model, but these different indices can lead to different conclusions (e.g. one supports the model, and another does not). We evaluated fit using two comparative fit indices (CFI and TLI) and two absolute fit indices (SRMS and RMSEA).

The fit for the hierarchical model should be interpreted with caution. As mentioned above, the sample size is low for accurate estimation. In addition, the comparative fit indices are artifactually limited and are not very informative because the RMSEA of the baseline model is less than 0.158 (Kenny, Kaniskan, & McCoach, 2015). Despite the complexity of the model and sample size limitations, the RMSEA value is at the recommended cutoff of .06, and the SRMR estimate is just above the recommended cutoff of .08 (Hu & Bentler, 1998, 1999). The model fit of the four quadrant models are adequate. All comparative fit indices are above .90, RMSEA values are at or just above .06, and all SRMR values are less than .08.

Fit Indices

Model	χ^2	DF	CFI	TLI	SRMR	RMSEA
Hierarchical	4573.12***	1062	0.80	0.79	0.09	0.06
Wisdom	195.91***	51	0.94	0.93	0.04	0.06
Humanity	221.28***	51	0.95	0.94	0.04	0.06
Courage	335.52***	51	0.92	0.90	0.06	0.08
Resilience	287.32***	51	0.95	0.93	0.04	0.07

*Note. ***p < .001*

Factor Loadings

All items in the hierarchical model and all four quadrant models had strong factor loadings. No items had loadings less than .40, which indicates that the items reflect the latent construct they were intended to measure. This provides additional evidence for internal structure validity.

RELIABILITY & AGREEMENT

Reliability is the degree to which an assessment is free of error, which is often operationalized as whether the assessment measures the same construct consistently (Crocker & Algina, 1986). There are different ways to evaluate the reliability of an assessment, including: test-retest, parallel forms, internal consistency, and inter-rater. Test-retest reliability shows the extent to which an assessment remains consistent over time, which is not relevant for a developmental feedback instrument. Ideally, leaders develop over time so ratings on a developmental assessment like the Positive Influence Predictor should NOT be consistent. Parallel forms reliability shows the consistency between two different versions of an assessment; however, it is also not relevant for the Positive Influence Predictor because there are no alternative forms. Internal consistency reliability is the degree to which items on a unidimensional scale are related. It provides evidence that items on the same scale are measuring the same construct. It is important for any type of assessment, and the results are discussed below.

Inter-rater reliability is the consistency of ratings between different raters, which is important but can be difficult to interpret for multi-rater assessments. It is assessed with correlation-based measures, which are biased if the variance of the scores is restricted. Multi-rater assessments are typically used for leaders in organizations who went through rigorous selection procedures, so it is unlikely that any will receive extremely low scores. This negatively skews the distribution of scores, which restricts variance and reduces estimates of inter-rater reliability (LeBreton, Burgess, Kaiser, Atchley, & James, 2003). Therefore, this information is supplemented with inter-rater agreement. Inter-rater reliability and agreement provide similar types of information, but agreement is not biased by restriction of variance.

Inter-rater agreement is the degree to which raters are interchangeable. It is a measure of absolute agreement, rather than the rank order consistency that is measured by inter-rater reliability (LeBreton & Senter, 2008). Because it is not measured with correlation-based indices, estimates of inter-rater agreement are not attenuated by the restriction of variance in multi-rater assessments. Inter-rater reliability and agreement results are presented together because they provide complementary information.

Internal Consistency

There are different statistics available for evaluating an assessment's internal consistency reliability. Split-half reliability is an internal consistency reliability statistic that is calculated by dividing an assessment in half and correlating responses from each half. However, this measure is limited because the different ways of splitting the assessment will yield different correlation coefficients.

Coefficient alpha (or Cronbach's alpha) is an alternative internal consistency statistic that corrects the problem of having many potential split-half reliability coefficients. Although it is rarely, if ever, calculated as such, coefficient alpha is mathematically equivalent to the arithmetic mean of all possible split-half reliabilities. The maximum value is 1.00, which indicates a perfect relationship between the items.

Although coefficient alpha is the most commonly reported reliability statistic, many psychometricians do not recommend its use (e.g., Sitjtsma, 2009; Trizano-Hermosilla & Alvarado, 2016; Yang & Green, 2011). The statistic has several assumptions that are unlikely to hold in practice (i.e., essential tau equivalence and uncorrelated errors), so researchers recommend using structural equation modeling estimates of reliability (Green & Yang, 2009). Therefore, we report both coefficient alpha, as is common practice, as well as a reliability index based on structural equation modeling – omega (McDonald, 1999).

Internal Consistency Reliability

Scale	Full Sample α	Independent Sample α	Independent Sample ω
Resilience			0.91
openness	0.76	0.76	0.77
inspiration	0.80	0.80	0.80
creativity	0.79	0.81	0.82
Courage			0.89
confidence	0.83	0.82	0.83
boldness	0.77	0.74	0.73
integrity	0.76	0.76	0.76
Humanity			0.89
likability	0.76	0.75	0.76
empathy	0.78	0.79	0.78
trust	0.69	0.69	0.70
Wisdom			0.84
perspective	0.69	0.69	0.70
focus	0.67	0.65	0.65
diligence	0.75	0.74	0.74

The table contains the estimates of coefficient alpha for the full sample and estimates of coefficient alpha and omega for the subset of the sample used for the factor analysis. The omega values were calculated using the four CFA models for each quadrant (see Internal Structure Validity Evidence section). Omega values at the quadrant level are also presented. The table shows that although three scales have at least one alpha estimate below 0.70, which is the general rule of thumb for reliability values, only the character strength of "focus" is slightly below this point using the recommended omega estimate. Internal consistency reflects the relationship between items on a scale, and when items are more related the reliability estimate increases. Considering the high factor loadings for the scale, indicating that all items are reflective of the construct they intended to measure (focus), good model fit, and bias of internal consistency estimates toward longer scales, no alterations to the "focus" scale were recommended based on the slightly low internal consistency estimate. However, if this scale does not improve with more data consideration of adding an item to this scale, rather than replacing an existing item, would be warranted.

Inter-rater Reliability

Intraclass correlation coefficients (ICCs) quantify the part of the variance in ratings attributed to differences between ratees. *ICC*(1) is the effect size estimate showing the effect of the ratee on observer ratings, and *ICC*(2) represents the reliability of the mean rating of the observers (LeBreton & Senter, 2008). *ICC*(1) values are interpreted like effect sizes, so a value of .10 would be medium, .25 would be large, and a small-medium effect of .05 provides evidence that warrants further consideration of inter-rater agreement (LeBreton & Senter, 2008). *ICC*(2) values are reliabilities of mean ratings, so values are interpreted such that values less than .40 are poor, between .40 and .75 are fair to good, and greater than .75 are excellent (Fleiss, 1986). The tables below with scale- and item-level *ICC*s show that all *ICC*(1) values are greater than .05, and most are greater than .10. In addition, all *ICC*(2) values are higher than .40.

Inter-rater Agreement

There are different statistics available for evaluating inter-rater agreement. In order to have a thorough analysis of inter-rater agreement we followed the recommendation of LeBreton and Senter (2008) and use multiple indices of inter-rater agreement. We chose to examine both a simple (average deviation) and more complex (r_{WG}) index. We calculated both at the scale- and item-level.

One of the most popular agreement indices is James, Demaree, and Wolf's (1993) r_{WG} and multi-item $r_{WG(J)}$. They measure proportional reduction in error variance by comparing the observed variance to the expected variance if there was no agreement between raters, which is represented by a null distribution. The most popular null distribution is the rectangular distribution, which represents a lack of agreement between raters as evenly spread across all possible values. This implies no bias in responding. Looking at the distribution of the Positive Influence Predictor data, there is evidence of a potential leniency bias. Even though raters are informed that their responses will be kept completely anonymous and will only be used for developmental purposes, there may still be a leniency bias if raters erroneously believe that their responses could somehow be identified, or if raters believe that low responses could hurt the ratee. Although a leniency bias would usually manifest as a skewed distribution, the middle of the scale is the best value on Positive Influence Predictor items, so a triangular distribution would represent leniency. Therefore, we also calculated *rwg* and $r_{WG(J)}$ with a null rectangular and triangular distribution using the expected error variance values provided by LeBreton and Senter (2008).

We also calculated the average deviation (Burke, Finkelstein, & Dusig 1999) for both the scales and items. Average deviation is a more practical index than r_{WG} in that it estimates agreement in the metric of the original scale. We estimated the average deviation around the median, rather than the mean, because it is a more robust test (Burke et al., 1999). Because the static describes distance from the central tendency, lower values of average deviation indicate higher levels of agreement.

The tables below show the results for scale- and item-level inter-rater agreement and reliability. The tables reflect the median values of each of the agreement statistics. Agreement statistics are calculated for every ratee, and the median is an unbiased summary

of those results. The typical cut off value for r_{WG} and $r_{WG(J)}$ is .70. All of the scale-level and almost all of the item-level r_{WG} values are above this cut off. Three median values using the more stringent triangular distribution fall just below .70 (.68, .69, .69). All average deviation values are well below the maximum recommended value of 1.5 for a 9-point scale (Burke & Dunlap, 2002). Overall this provides excellent support for inter-rater agreement.

Median Inter-Rater Reliability & Agreement of Scales

Scale	$r_{WG(J)}$ Rectangular	$r_{WG(J)}$ Triangular	Average Deviation	*ICC*(1)	*ICC*(2)
openness	0.97	0.93	0.53	0.14	0.60
inspiration	0.97	0.91	0.61	0.21	0.71
creativity	0.97	0.93	0.56	0.13	0.58
confidence	0.97	0.91	0.61	0.28	0.78
boldness	0.97	0.92	0.57	0.20	0.70
integrity	0.99	0.97	0.25	0.18	0.67
likability	0.98	0.94	0.45	0.24	0.74
empathy	0.97	0.93	0.5	0.23	0.74
trust	0.98	0.96	0.38	0.20	0.70
perspective	0.97	0.92	0.54	0.14	0.59
focus	0.97	0.93	0.52	0.17	0.66
diligence	0.97	0.93	0.5	0.20	0.70

Median Inter-Rater Reliability & Agreement of Items

Item	r_{WG} Rectangular	r_{WG} Triangular	Average Deviation	*ICC*(1)	*ICC*(2)
inspiration 1	0.9	0.77	0.57	0.16	0.63
inspiration 2	0.86	0.7	0.67	0.15	0.61
inspiration 3	0.9	0.77	0.57	0.22	0.72
inspiration 4	0.88	0.73	0.64	0.14	0.59
creativity 1	0.92	0.82	0.46	0.10	0.48
creativity 2	0.89	0.75	0.62	0.11	0.51
creativity 3	0.91	0.8	0.5	0.08	0.42
creativity 4	0.86	0.69	0.67	0.13	0.56
confidence 1	0.86	0.68	0.69	0.25	0.75
confidence 2	0.89	0.76	0.58	0.17	0.65
confidence 3	0.9	0.77	0.58	0.22	0.72
confidence 4	0.89	0.75	0.6	0.16	0.64
boldness 1	0.86	0.69	0.67	0.16	0.62
boldness 2	0.88	0.73	0.64	0.16	0.63
boldness 3	0.92	0.82	0.5	0.14	0.58
boldness 4	0.91	0.79	0.5	0.18	0.66
integrity 1	0.96	0.91	0.28	0.10	0.50
integrity 2	0.95	0.89	0.3	0.13	0.57
integrity 3	0.96	0.91	0.25	0.12	0.54
integrity 4	0.97	0.94	0.18	0.16	0.63
diligence 1	0.88	0.73	0.6	0.16	0.63
diligence 2	0.89	0.75	0.57	0.18	0.66
diligence 3	0.94	0.86	0.4	0.13	0.57
diligence 4	0.91	0.81	0.5	0.13	0.58
focus 1	0.9	0.78	0.53	0.10	0.49
focus 2	0.89	0.75	0.58	0.11	0.52
focus 3	0.92	0.82	0.5	0.09	0.49
focus 4	0.91	0.8	0.5	0.13	0.57
perspective 1	0.89	0.75	0.59	0.10	0.50
perspective 2	0.9	0.78	0.54	0.10	0.52
perspective 3	0.9	0.77	0.53	0.10	0.51
perspective 4	0.92	0.81	0.47	0.09	0.46
trust 1	0.95	0.9	0.31	0.16	0.63
trust 2	0.95	0.89	0.33	0.12	0.56
trust 3	0.95	0.89	0.33	0.12	0.55
trust 4	0.91	0.8	0.5	0.20	0.69
empathy 1	0.87	0.72	0.62	0.17	0.65
empathy 2	0.92	0.82	0.45	0.14	0.60
empathy 3	0.9	0.77	0.56	0.16	0.60
empathy 4	0.95	0.89	0.33	0.18	0.67
likability 1	0.92	0.83	0.44	0.13	0.56
likability 2	0.94	0.86	0.38	0.14	0.60
likability 3	0.91	0.79	0.5	0.20	0.70
likability 4	0.9	0.77	0.5	0.19	0.69
openness 1	0.9	0.78	0.54	0.13	0.57
openness 2	0.9	0.77	0.58	0.09	0.47
openness 3	0.9	0.78	0.56	0.12	0.56
openness 4	0.92	0.82	0.47	0.08	0.44

CONCLUSIONS

The present evaluation provided evidence that the revised version of the Positive Influence Predictor has sound psychometric properties. Analyses indicated that all but one of the 12 unidimensional subscales showed acceptable omega values. This shows that items on the subscales are related, which reflects high reliability. *ICC* values lend evidence to inter-rater reliability for all items and scales. In addition, all 12 subscales had a high degree of inter-rater agreement. All median $r_{WG(J)}$ values exceeded the minimum cutoff, even using the more conservative triangular null distributions. All items r_{WG} estimates exceeded .70 using the rectangular distribution as the null, and most items also exceeded this cutoff with the more conservative triangular distribution as the null. All items and scales demonstrated high agreement using average deviation.

The Positive Influence Predictor also showed strong validity evidence based on internal structure. A second-order confirmatory factor analysis provided some evidence that the overall factor structure of the Positive Influence Predictor fit the implied structure of the Tilt Framework, but inferences from this model are limited by the small sample size. Four supplemental quadrant-level models showed acceptable fit to the *T*ilt model. Factor loadings were high for all items in all five models.

REFERENCES

Burke, M. J., & Dunlap, W. P. (2002). Estimating interrater agreement with the average deviation index: A user's guide. *Organizational Research Methods, 5*(2), 159-172. doi:10.1177/1094428102005002002

Burke, M. J., Finkelstein, L. M., & Dusig, M. S. (1999). On average deviation indices for estimating interrater agreement. *Organizational Research Methods, 2*(1), 49-68. doi:10.1177/109442819921004

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology, 66*, 4-19. doi:10.1016/j.jesp.2015.07.006

Fleiss, J. (1986). *The Design and Analysis of Clinical Experiments*. Wiley, New York.

Francavilla, N. M., Meade, A. W., & Young, A. L. (2018). Social interaction and internet-based surveys: Examining the effects of virtual and in-person proctors on careless response: Proctors and careless response. *Applied Psychology,* doi:10.1111/apps.12159

Green, S. B., & Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika, 74*, 121-135. doi:10.1007/s11336-008-9098-4

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*(4), 424-453. doi:10.1037/1082-989X.3.4.424

Hu, L., & 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. http://dx.doi.org/10.1080/10705519909540118

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods Research, 44*(3), 486–507. doi:10.1177/0049124114543236

Kline, R. B. (2016). *Principles and practice of structural equation modeling (4th ed.)*. New York, NY, US: Guilford Press.

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Erlbaum.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*, 437–455. doi:10.1037/a0028085

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of cronbach's alpha. *Psychometrika, 74*, 107-120. doi:10.1007/s11336-008-9101-0

Trizano-Hermosilla, I., & Alvarado, J. M. (2016). Best alternatives to cronbach's alpha reliability in realistic conditions: Congeneric and asymmetrical measurements. *Frontiers in Psychology, 7,* 769-769. doi:10.3389/fpsyg.2016.00769

Yang, Y., & Green, S. B. (2011). Coefficient alpha: A reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment, 29*, 377-392. doi:10.1177/0734282911406668