A Validation of the Non-Parametric Continuous Norming Procedure

Melissa Jane Siy[1] and Francisco N. de los Reyes[2]
[1]Philippine Survey and Research Center
[2]School of Statistics, University of the Philippines Diliman

This study analyzed and validated the statistical aspect of the non-parametric continuous norming technique, which is a method used in creating scores in psychometric tests. Using the Work Profile Questionnaire - Emotional Intelligence (WPQei) with Filipino sample respondents, the study was able to demonstrate how the norming technique can be used to create age-group-based scores (age norms). Based on the results, the models from the technique can produce useable scores in practice with acceptable adjusted R-squared values; however, some challenges emerged with regard to the process of choosing the smoothing parameter, consistency of the significant variables and coefficient signs of the model, and the calculation of the score tables. Bootstrapping is recommended in improving the robustness of the technique.

Keywords: norming, WPQei, age-group-based score tables, bootstrap

1. Introduction

Psychometric or psychological tests are used in diverse settings and are made to measure intelligence, aptitude, or personality traits. While psychometric tests play a vital role in the decision-making process of many test users, the results can likewise be life-changing to the recipients. Hence, the development and use of these tests should be given serious attention.

There are many procedures involved in test development, but the study intends to focus mainly on the establishment of test norms, or simply known as norming. Norms are the test performance of a standardization or normative sample, which is assumed to be representative of a target population (Anastasi & Urbina, 2009). Given the norms, an individual's score can be compared with a reference group's score distribution; thus, the relative standing of an individual can be assessed. Ideally, the test-taker and the normative sample should have similar characteristics in order for the interpretation to be sound. In the Philippines, although there are local tests, international tests are more dominant. As expected, normative samples primarily consist of American and European middle class, leading to a lack of representation for other ethnicities (Groth-Marnat, 2003). The issue of inappropriate normative sample is one component of the test bias problem

THE_PHILIPPINE_STATISTICIAN_Vol_69_No_2_2020_6x9.indd 17 2/19/2021 10:10:35 AM

that has been known for decades in the field of assessment. Although there are tests that have noted this concern, it is still common for practitioners to encourage the creation of own norms so that the normative sample is more localized, current, and less biased (Fischer & Milfont, 2010; Groth-Marnat, 2003; Reynolds & Ramsay, 2003). This is especially helpful for countries like the Philippines that do not have many local tests.

The study extends from the work of Lehnard, Lehnard, Sebastian, and Sergerer (2016), wherein a novel norming approach was introduced. It is thus the objective of this study to open the discussion about the derivation of norm scores, to contribute statistical and practical insights for the improvement of these techniques, and to encourage Filipino test users and test developers to overcome the barriers in creating own norms and find time to understand and develop these methods. Specifically, the study seeks a good-fitting model that is also optimal from a practical perspective, produces usable age norms and comparable to a reference norm table currently used in practice.

2. Operational Framework, Scope and Limitation

To give a better visualization of the objectives, Figure 1 pieces together all the tasks, the processes or methods involved, and the expected output from which all inferences were made.

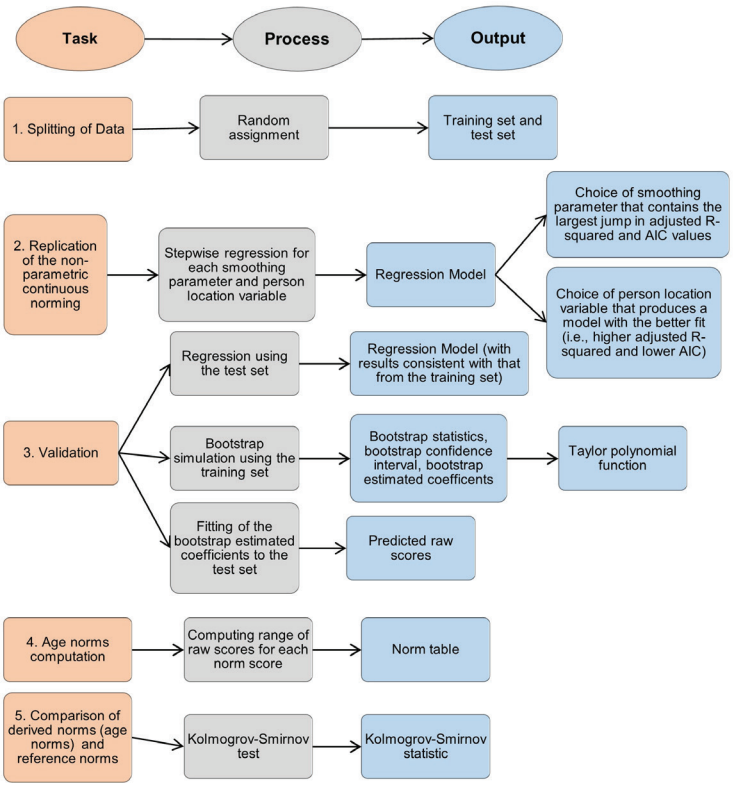


Figure 1. Operational Framework

THE_PHILIPPINE_STATISTICIAN_Vol_69_No_2_2020_6x9.indd 18 2/19/2021 10:10:36 AM

Given that norming encompasses a wide range of procedures, the study limits its scope to the derivation of norm scores only. Sampling is not included in the scope of the study framework. Observations were a simple collection of test takers over the period of time from November 2016 to March 2017. As such, sampling errors and measurement errors are assumed part of the overall statistical errors in all inferential aspects of this work. Since the study investigates a norming technique wherein a continuous covariate is involved, age was chosen as the covariate because of the evidence in literature about its relation with emotional intelligence.

3. Methodology

Test Instrument. The data was accessed from a local testing company in the Philippines. Instead of using an aptitude test as in Lehnard et al. (2016), the personality test was used. This is the Work Profile Questionnaire – Emotional Intelligence (WPQei) (Cameron, 2004) which is an 84-item questionnaire designed to measure one's ability to regulate a person's emotions in the work setting. WPQei is answered using a Likert scale; one item in the test has a minimum raw score of 1 and a maximum raw score of 5. Each of the seven factors consists of 12 items. This instrument measures seven factors summarized below, along with their respective internal consistency statistic (Cronbach's Alpha):

1. Innovation refers to one's problem solving style. Low scorers are more inclined to be rule-implementers, while high scorers lean towards risk-taking and novelty (Cronbach's Alpha = 0.70).
2. Self-Awareness measures one's awareness and ability to understand his/her personal strengths and weaknesses (Cronbach's Alpha = 0.65).
3. Intuition is the extent to which one is able to rely on his instinct and feelings when thinking or making decisions (Cronbach's Alpha = 0.66).
4. Emotions refers to one's ability to identify and control his/her emotional responses (Cronbach's Alpha = 0.68).
5. Motivation measures one's drive and work commitment (Cronbach's Alpha = 0.76).
6. Empathy pertains to one's ability to understand other people's opinions and feelings (Cronbach's Alpha = 0.76).
7. Social Skills refers to one's ability to handle social situations (Cronbach's Alpha = 0.76).

Norms. The WPQei makes use of the sten (standard ten) scale as its norm score. Stens have a mean of 5.5 and standard deviation of 2 and are only reported as whole numbers, ranging from 1 to 10. By using stens, the normal distribution is divided into ten regions, wherein sten scores from 4 to 7 are expected to be the "average", containing 68% of the sample. Stens 2 to 9 are equivalent to one half of a z-score each. Meanwhile, sten 1 contains all scores below -2 z-score and sten 10 covers all scores above +2 z-score. Stens are computed using the following formula (Neukrug & Fawcett, 2014):

$$Sten = z - score(standard\backslash deviation) + mean = z - score(2) + 5.5.$$

THE_PHILIPPINE_STATISTICIAN_Vol_69_No_2_2020_6x9.indd 19 2/19/2021 10:10:36 AM

The current WPQei documentation only describes general managerial and professional norms, gender-based norms, and norms for degree-holders. This study adds to this collection by creating age-based norms using the non-parametric continuous norming of Lehnard et al. (2016).

Age is an important factor in emotional intelligence. In fact, emotional intelligence was shown to meet the classical criteria for intelligence; one of which is that it develops with age and experience (Mayer, Caruso & Salovey, 1999). Kafetsios (2004) was able to show in his research that older individuals had higher emotional intelligence scores, but only in three out of four scales of the Mayer, Salovey, and Caruso emotional intelligence test (MSCEIT V2.0). Given these studies and its practical relevance, age was chosen as the continuous covariate.

Sample. Respondents answered the WPQei from November 2016 to March 2017. The sample size was 1,117 test takers. All observations with item non-response were excluded, along with one respondent who reported having an age of 72. From the entire data set, approximately 30% (330 test takers) was randomly selected and assigned as the test set, the remaining 70% (787 test takers) was assigned as the training set. Statistical properties of the entire data set were compared with the training and test set to ensure consistency, particularly in skewness. Skewness was measured using the D' Agostino Skewness test (Komsta & Novomestky, 2015). The whole data consists of respondents with ages ranging from 15 to 59 years old. The mean age is 25.095 (SD = 7.349). Both sexes were fairly represented with 50.8% of the data consisting of females, while the rest were males. The training set consists of respondents with mean age of 25.267 (SD = 7.372), consisting of 50.2% females. While the test set comprises of respondents with mean age of 24.685 (SD = 7.286) and 52.1% females.

Non-parametric Continuous Norming. The first part of the analysis is the replication of the nonparametric continuous norming method of Lehnard et al. (2016). In this method, the raw score is modeled as a continuous function of an explanatory variable (i.e., age) and the person's location in the distribution (e.g., percentile or normalized standard score/norm score). Taylor polynomials are used to approximate the function, which is expressed as:

$$r(l,a) = \sum_{s,t=0}^{k} c_{st} l^s a^t$$

where r = raw score
l = location (percentile or normalized standard score)
a = age
k = smoothing parameter, and
c_{st} = constants determined through regression.

This was derived from the following Taylor series centered at point $P(l_0, a_0)$, expressed as:

$$r(l,a) = \sum_{s,t=0}^{\infty} \frac{1}{s!\,t!} \frac{\partial^{s+t} f(l_0, a_0)}{\partial l^s \partial a^t} (l - l_0)^s (a - a_0)^t$$

THE_PHILIPPINE_STATISTICIAN_Vol_69_No_2_2020_6x9.indd 20 2/19/2021 10:10:36 AM

Following are the step-by-step procedures used in replicating the non-parametric continuous norming method:

1. The training set was arbitrarily divided into the five age (*a*) brackets: Teens (15-19 years old), 20s (20-29 years old), 30s (30-39 years old), 40s (40-49 years old), and 50s (50-59 years old).
2. Person location variables (*l*), normalized standard scores (i.e., stens) and percentiles, were computed per age bracket, followed by the powers of *a*, *l*, and the products of these powers. These powers served as the explanatory variables in the regression procedures. Table 1 summarizes the explanatory variables up to smoothing parameter $k = 2$. In the study, powers up to $k = 9$ were investigated.
3. Table 1. **Summary of Explanatory Variables per Smoothing Parameter (k)**

k	Explanatory Variables
1	a, l, la
2	a, a^2, l, l^2, la, l^2a, la^2, l^2a^2

    a. The norming procedure allows for either normalized standard scores or percentiles as the person location variable; however, in the study, both the normalized standard scores and percentiles were used in separate runs of the procedure in order to observe any differences that may result from using different kinds of person locations.
    b. In practice, it is not necessary for the practitioner to check all 9 smoothing parameters. In fact, Lehnard, et al. (2016) recommends $\boldsymbol{k = 5}$ as the starting point, but the choice of the appropriate smoothing parameter can vary.
4. For each smoothing parameter and person location variable, stepwise multiple regression was employed using the raw score as the dependent variable and the powers and products of the powers as the explanatory variables. The F-test selection criterion was used in selecting the final model.
    a. For each smoothing parameter, the adjusted R-squared and the AIC of the final model were reported. These were used in determining the person location variable and smoothing parameter to be used in the subsequent validation procedures. The person location was chosen based on which produced better fit models, while the smoothing parameter was chosen based on the largest jump in the adjusted R-squared and AIC value. The largest jump in value would mean that smoothing parameters above the chosen point would not contribute much difference in terms of the adjusted R-squared and AIC value, thus the lower smoothing parameter would suffice.

THE_PHILIPPINE_STATISTICIAN_Vol_69_No_2_2020_6x9.indd 21 2/19/2021 10:10:36 AM

b. Originally, the Taylor polynomial function can already be defined from the stepwise regression by incorporating the significant variables and their respective coefficients as the constants. However, in the present study, the Taylor polynomial functions were defined after the final models underwent validation analyses. Procedures in the validation are discussed in the next section.

Validation. There were three main procedures in the validation analyses. First was the comparison of the training set and test set regression results. Having chosen the smoothing parameter and person location variable from the replication, the variables of the final model in the training set were entered in the test set. Afterwards, the test set and training set regression results were compared, specifically whether they produced the same significant variables and coefficient signs.

The second validation procedure involved bootstrapping. Bootstrap simulation was employed as another method to check the consistency of the regression results. One thousand (1,000) bootstrap samples of size 787 were generated from the training set. Each bootstrap sample underwent regression based on the final model of the training set regression. From there, the bootstrap statistics, bootstrap estimated coefficients, and bootstrap confidence intervals were computed. The outputs present the following:

1. ***bootstrap estimated coefficient***, which is the arithmetic mean of the regression coefficients produced from the 1000 bootstrap samples;
2. ***bias,*** which is the difference between bootstrap estimated coefficient and the original estimate from the training set; and
3. ***bootstrap confidence interval,*** which is a two-sided non-parametric interval computed using the basic bootstrap confidence interval method (Davison & Hinkley, 1997):

$$\left[2\hat{\theta} - \hat{\theta}^*_{1-\frac{\alpha}{2}}, 2\hat{\theta} - \hat{\theta}^*_{\frac{\alpha}{2}}\right]$$

where $\hat{\theta}$ is the sample estimate, $\hat{\theta}^*$ is the bootstrap estimate, α is the significance level, and $\hat{\theta}^*_\alpha$ is the α-quantile of the bootstrap distribution. The significant variables and the signs of the bootstrap estimated regression coefficients were compared with the training set regressions results to check for consistency. The Taylor polynomial function was then defined using significant variables from the bootstrap simulation, along with the estimated bootstrap coefficients as the constants. This particular step is different from Lehnard et al. (2016) wherein the Taylor polynomial function was defined from the results of the stepwise regression results. Third, the function was fitted to the test set and the predicted raw scores and their descriptive properties were checked. Specifically, the predicted raw scores are expected to be within the expected minimum and maximum test raw score. This would verify whether the model is useable in practice.

THE_PHILIPPINE_STATISTICIAN_Vol_69_No_2_2020_6x9.indd 22 2/19/2021 10:10:36 AM

Norm Tables. The Taylor polynomial function from the validation was used in creating the norm tables. Norm scores are computed by inserting the lower bound of the person location and the mean age of the age bracket in the Taylor polynomial function. The same was done for the upper bound of the person location. As a result, every person location value had a corresponding range of raw scores.

Kolmogorov-Smirnov Goodness-of-Fit Test. The respondents' sten score distribution using the reference norms and the sten score distribution using the derived norms/age norms were compared to check whether the distributions are statistically different. Inferences were based on the Kolmogorov-Smirnov statistic (*D*). The null hypothesis tested was H_0: The derived norm scores and the reference norm scores are the same distribution, versus the alternative hypothesis H_a: The derived norm scores and the reference norm scores are different distributions. H_0 is rejected if the *D* is greater than the critical value at significance level 0.05.

4. Results and Discussion

Table 2 compares the skewness of the entire data, the training set, and the test set, respectively. The p-values in the tables are from the D'Agostino Skewness Test. Based on the entire data, Empathy is negatively skewed, while Innovation and Intuition are positively skewed. The rest of the factors are normally distributed. The same results were observed in the training set and the test set.

Table 2. Descriptive Statistics

Factor	Entire Data		Training Set		Test Set	
	Skewness	Std. Error	Skewness	Std. Error	Skewness	Std. Error
Emotions	.140	0.073	.113	0.087	.207	0.134
Empathy	-.453*	0.073	-.429*	0.087	-.504*	0.134
Innovation	.500*	0.073	.360*	0.087	.824*	0.134
Intuition	.283*	0.073	.208*	0.087	.461*	0.134
Motivation	.013	0.073	-.023	0.087	.100	0.134
Self-Awareness	-.018	0.073	-.017	0.087	-.023	0.134
Social Skills	.028	0.073	.015	0.087	.061	0.134
Note:	n = 1117; p<0.05*		n = 787; p<0.05*		n = 330; p<0.05*	

Non-Parametric Continuous Norming. After executing the norming technique for each of the 9 smoothing parameters, the adjusted R-squared and AIC of the final models from the stepwise regression (using either sten or percentile as the person location variable) were compared. Figure 2 displays a graph of the adjusted R-squared and AIC for stens and percentiles for each smoothing parameter of each factor. Starting with the Emotions factor, results show that across all smoothing parameters, stens produced final models with better fit (i.e., higher adjusted R-squared and lower AICs) compared to percentiles, but the

THE_PHILIPPINE_STATISTICIAN_Vol_69_No_2_2020_6x9.indd 23 2/19/2021 10:10:36 AM

difference of adjacent adjusted R-squared of stens and percentiles decreased as the smoothing parameter increased. The same occurred in the AICs of the stens and percentiles. A similar trend was observed for all the other factors, Empathy, Innovation, Intuition, Motivation, Self-Awareness, and Social Skills. As a result, all factors used stens, instead of percentiles, as the person location variable for the subsequent validation analyses. Although Lehnard et al. (2016) allowed the use of percentiles as the person location variable, normalized standard scores, or in this case, stens, were the preferred variable because better fit models were produced.

As for the optimal smoothing parameter, all factors produced the largest jump from $k = 1$ to $k = 2$ for stens. Beyond $k = 2$, changes in adjacent adjusted R-squared and AICs were smaller. Moreover, the adjusted R-squared and the AIC started to fluctuate as the smoothing parameter increased. For percentiles, the largest jump was from $k = 2$ to $k = 3$, one point larger than the stens. As such, all $k = 2$ was the chosen smoothing parameter for all factors. Table 3 summarizes the adjusted R^2 and the AIC of the chosen final models for each factor.

Table 3. Adjusted R^2 and AIC of the Final Models for k=2 (Smoothing Parameter)

	Adjusted R-squared	AIC
Emotions	0.966	-25.742
Empathy	0.960	353.366
Innovation	0.963	-122.758
Intuition	0.980	-342.746
Motivation	0.957	434.701
Self-Awareness	0.965	294.237
Social Skills	0.973	42.782

Validation. The variables from the final model of the training set were entered to the test set. Table 4 shows the side-by-side comparison of the results using the training set and the test set. Across all factors the adjusted R-squared did not drastically change; however, in terms of consistency, only Empathy was consistent in terms of the significant variables and the sign of the coefficients. This inconsistency is a disadvantage for the norming procedure since the significant variables and the signs of the coefficients determine the Taylor polynomial function, and eventually the norm scores. The results reflect that the norming method can produce different norms scores from sample to sample.

As a follow-up in the validation process, 1000 bootstrap samples of size 787 were generated from the training set. Each bootstrap sample underwent regression and produced its own set of coefficients. The output, seen in Table 5, reveal that the bootstrap results and the training set regression results produced the same significant variables. Moreover, coefficient signs were the same. This was a

THE_PHILIPPINE_STATISTICIAN_Vol_69_No_2_2020_6x9.indd 24 2/19/2021 10:10:36 AM

consistent trend for all factors, except for Social Skills. The training set regression results and the bootstrap results for this factor did not produce the same significant variables. However, the bootstrap estimated coefficient values were close to the training set regression coefficients.

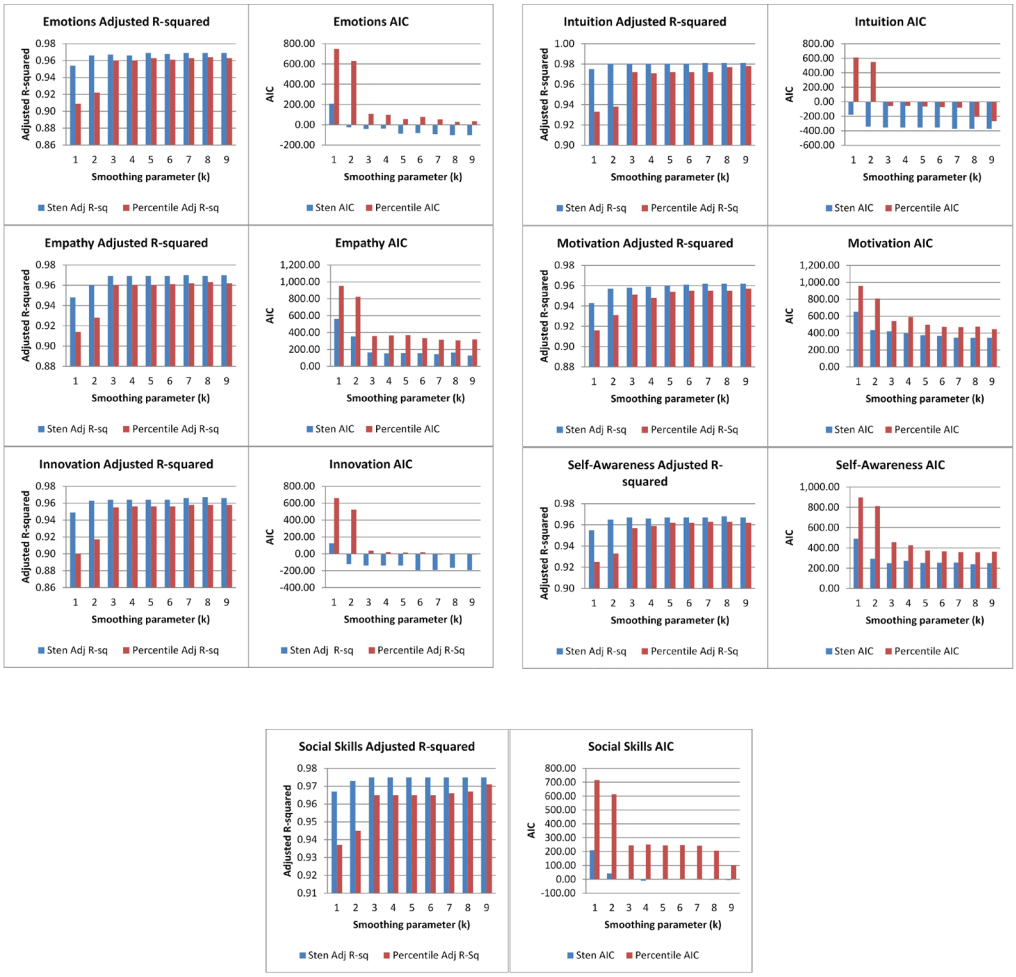


Figure 2. Adjusted R-squared and AIC of Final Models for the Instrument Factors

THE_PHILIPPINE_STATISTICIAN_Vol_69_No_2_2020_6x9.indd 25 2/19/2021 10:10:37 AM

Table 4. Comparison of Final Model Statistics Between Training and Test Sets by Factor

Emotions	Training Set				Test Set			
	Coef	Std. Error	Adj R^2	AIC	Coef	Std. Error	Adj R^2	AIC
Cons	7.372*	.468			15.333*	.595		
l1	2.239*	.076			2.086*	.104		
a1	.593*	.029	0.966	-25.742	.066	.036	0.975	-110.208
a2	-.007*	4.46E-04			1.377E-04	.001		
l2	.027*	.007			.054*	.054		
Empathy								
Cons	8.419*	.602			10.107*	.740		
l1	4.190*	.099			4.957*	.127		
a1	.531*	.037	0.960	353.366	.286*	.044	0.976	18.711
l2	-.107*	.009			-.155*	.011		
a2	-.006*	.001			-.003*	.001		
Innovation								
Cons	15.130*	.235			18.141*	.458		
l1	.370*	.098			.210	.193		
l1a1	.086*	.005	0.963	-122.758	.036*	.008	0.944	98.543
l1a2	-.001*	8.00E-05*			-4.772E-04*	1.39E-04		
l2	.053*	.006			.144*	.012		
a2	.001*	2.14E-04*			-3.607E-04*	4.26E-04		
Intuition								
Cons	10.786*	.385			18.012*	.715		
l1	2.163*	.061			1.621*	.123		
a1	.369*	.023	0.980	-342.756	.013	.043	0.970	6.440
l2	.053*	.005			.115*	.011		
a2	-.003*	3.64E-04			1.683E-04	.001		
Motivation								
Cons	11.950*	.621			18.047*	.776		
a1	.432*	.041			.058	.050		
a2	-.009*	.001	0.957	434.701	-4.057E-03*	.001	0.969	75.184
l2	.248*	.008			.248*	.011		
l1a1	.130*	.004			.121*	.006		
l2a1	-.011*	4.44E-04*			-9.445E-03*	.001		
Self-Awareness								
Cons	13.945*	.563			17.868*	.714		
a1	.331*	.038			.043	.046		
a2	-.007*	.001	0.965	294.237	-4.825E-03*	.001	0.976	0.588
l1a1	.123*	.004			.139*	.005		
l2	.257*	.007			.276*	.009		
l2a1	-.010*	4.13E-04*			-.012*	.001		
Social Skills								
Cons	6.622*	.570			9.042*	.821		
l1	2.508*	.149			2.918*	.231		
a1	.558*	.030			.258*	.040		
a2	-.009*	.001	0.973	42.482	-1.195E-03	.001	0.978	-40.393
l1a2	.001*	1.91E-04*			-1.083E-04	3.43E-04		
l2a2	-6.716E-05*	1.70E-05*			-1.744E-05	3.00E-05		
l2	.039*	.013			.038	.020		
	Note: n= 787; k = 2; l = sten; a = age; p<0.05*				Note: n= 330; k = 2; l = sten; a = age; p<0.05*			

THE_PHILIPPINE_STATISTICIAN_Vol_69_No_2_2020_6x9.indd 26 2/19/2021 10:10:37 AM

Table 5. Bootstrap Validation Results

Emotions	Original	Bias	Standard Error	Bootstrap Confidence Interval	Bootstrap Estimated Coefficients
Cons	7.372	0.014	0.567	(6.271, 8.551)	7.386*
l1	2.239	-4.574E-03	0.087	(2.074, 2.419)	2.234*
a1	0.593	1.927E-04	0.037	(0.513, 0.664)	0.593*
a2	-7.123E-03	-9.058E-06	5.893E-04	(-8.218E-03, -5.877E-03)	-7.132E-03*
l2	0.027	4.202E-04	7.414E-03	(0.011, 0.041)	0.027*
Empathy					
Cons	8.419	-0.090	0.968	(6.736, 10.500)	8.328*
l1	4.190	1.530E-03	0.090	(4.010, 4.354)	4.192*
a1	0.531	6.308E-03	0.067	(0.388, 0.651)	0.537*
l2	-0.107	-2.322E-04	7.848E-03	(-0.122, -0.091)	-0.107*
a2	-5.924E-03	-9.934E-05	1.100E-03	(-7.861E-03, -3.550E-03)	-6.023E-03*
Innovation					
Cons	15.130	0.024	0.310	(14.478, 15.680)	15.153*
l1	0.370	-0.016	0.122	(0.158, 0.632)	0.354*
l1a1	0.086	6.277E-04	6.446E-03	(0.072, 0.098)	0.086*
l1a2	-1.208E-03	-1.057E-05	1.286E-04	(-1.427E-03, -9.312E-04)	-1.219E-03*
l2	0.053	5.511E-04	8.166E-03	(0.036, 0.068)	0.053*
a2	1.162E-03	-1.329E-06	3.260E-04	(5.196E-04, 1.820E-03)	1.161E-03*
Intuition					
Cons	10.786	0.029	0.468	(9.858, 11.669)	10.815*
l1	2.163	-3.742E-03	0.073	(2.022, 2.311)	2.160*
a1	0.369	-1.334E-03	0.030	(0.311, 0.430)	0.367*
l2	0.053	3.774E-04	6.792E-03	(0.039, 0.066)	0.053*
a2	-3.276E-03	1.788E-05	4.707E-04	(-4.175E-03, -2.380E-03)	-3.258E-03*
Motivation					
Cons	11.950	-0.115	0.800	(10.546, 13.659)	11.835*
a1	0.432	7.912E-03	0.054	(0.317, 0.525)	0.440
a2	-8.933E-03	-1.284E-04	7.955E-04	(-1.028E-02, -7.087E-03)	-9.061E-03*
l2	0.248	5.856E-04	0.010	(0.227, 0.266)	0.248*
l1a1	0.130	4.365E-05	4.335E-03	(0.121, 0.139)	0.130*
l2a1	-0.011	-2.625E-05	5.163E-04	(-0.012, -0.010)	-0.011*
Self-Awareness					
Cons	13.945	-0.081	0.758	(12.508, 15.550)	13.864*
a1	0.331	5.266E-03	0.054	(0.217, 0.428)	0.336*
a2	-7.290E-03	-9.105E-05	7.694E-04	(-8.758E-03, -5.685E-03)	-7.381E-03*
l1a1	0.123	1.867E-04	4.895E-03	(0.113, 0.132)	0.123*
l2	0.257	5.227E-04	8.743E-03	(0.239, 0.274)	0.257*
l2a1	-0.010	-3.709E-05	4.535E-04	(-1.118E-02, -9.337E-03)	-0.010*
Social Skills					
Cons	6.622	0.023	0.817	(4.973, 8.307)	6.645*
l1	2.508	-0.025	0.241	(2.090, 3.047)	2.482*
a1	0.558	3.281E-03	0.044	(0.466, 0.639)	0.561*
a2	-8.747E-03	-1.774E-04	1.389E-03	(-1.088E-02, -5.437E-03)	-8.925E-03*
l1a2	8.552E-04	4.450E-05	3.690E-04	(-2.169E-05, 1.443E-03)	8.997E-04
l2a2	-6.716E-05	-3.532E-06	3.119E-05	(-1.204E-04, 4.504E-06)	-7.069E-05
l2	0.039	2.051E-03	0.022	(-8.014E-03, 7.622E-02)	0.041

THE_PHILIPPINE_STATISTICIAN_Vol_69_No_2_2020_6x9.indd 27 2/19/2021 10:10:38 AM

The Taylor polynomial per factor (shown below) was then defined from the bootstrap results using the significant variables and their respective bootstrap estimated coefficients as the constants. These functions were then fitted to the test set.

$$\widehat{emotions\ raw}\ score\ k2 = 7.386 + (2.234 * l1) + (0.593 * a1) + (-7.132 \times 10^{-3} * a2) + (0.027 * l2)$$

$$\widehat{empathy\ raw}\ score\ k2 = 8.328 + (4.192 * l1) + (0.537 * a1) + (-0.10 * l2) + (-6.023 \times 10^{-3} * a2)$$

$$\widehat{innovation\ raw}\ score\ k2 = 15.153 + (0.354 * l1) + (0.086 * l1a1) + (-1.219 \times 10^{-3} * l1a2) + (0.053 * l2) + (1.161 \times 10^{-3} * a2)$$

$$\widehat{intuition\ raw}\ score\ k2 = 10.815 + (2.160 * l1) + (0.367 * a1) + (0.053 * l2) + (3.258 \times 10^{-3} * a2)$$

$$\widehat{motivation\ raw}\ score\ k2 = 11.835 + (0.440 * a1) + (-9.061 \times 10^{-3} * a2) + (0.248 * l2) + (0.130 * l1a1) + (-0.011 * l2a1)$$

$$\widehat{self-awareness\ raw}\ score\ k2 = 13.864 + (0.336 * a1) + (-7.831 \times 10^{-3} * a2) + (0.123 * l1a1) + (0.257 * l2) + (-0.010 * l2a1)$$

$$\widehat{social\ skills\ raw}\ score\ k2 = 6.645 + (2.842 * l1) + (0.561 * a1) + (-8.925 \times 10^{-3} * a2)$$

The resulting predicted raw scores were discovered to be within the expected raw score range. Thus, the model, using k = 2, is useable in practice.

Norm Tables. Age norms, found in Table 6, were computed using Lehnard et al.'s (2016) recommended procedure. Each sten in the table has a corresponding range of raw scores. The only issue is that the lower/upper-bound were repeated in adjacent stens. In practice, the choice of the lower/upper bound can be determined by the practitioner; nonetheless, a rule of thumb regarding this issue should have been established. For the purpose of the study, the repeating raw scores were used in the higher sten. For example in the Emotions factor for teens, sten one would contain 18 and below, sten two 19 to 20, sten three 21 to 23, and so on. An additional issue in the Motivation factor emerged in sten 10 of the fifties age group, wherein the raw score upper bound and lower bound differed only by decimal values. Since raw scores are only whole numbers, this led to a single-digit raw score in sten 10. This encounter was another hurdle of the norming technique. A higher smoothing parameter may need to be considered if similar situations are encountered in practice. However, in this case, the norms are still useable. In such a way that, raw score 43 can be assigned to sten 9 and raw score 44 to sten 10. For the purposes of the study, the repeating raw scores were used in the higher sten as in previous factors. In practice, stens 1 and 10 will accommodate the minimum raw score and the maximum raw score, respectively, if these values did not result from the Taylor polynomial computation. Since the reference norms from the manual do not contain age norms, Table 6 can be used to assess the test takers' emotional intelligence scores based on their age group.

THE_PHILIPPINE_STATISTICIAN_Vol_69_No_2_2020_6x9.indd 28 2/19/2021 10:10:38 AM

Kolmogrov-Smirnov Test. Results from the Kolmogrov-Smirnov test (See Table 7) showed that the distribution of the reference norms (as provided in the test manual) and the derived norms/age norms (using the non-parametric continuous norming) are statistically different from each other. Thus, these norms cannot be used interchangeably.

Table 6. Age Norms

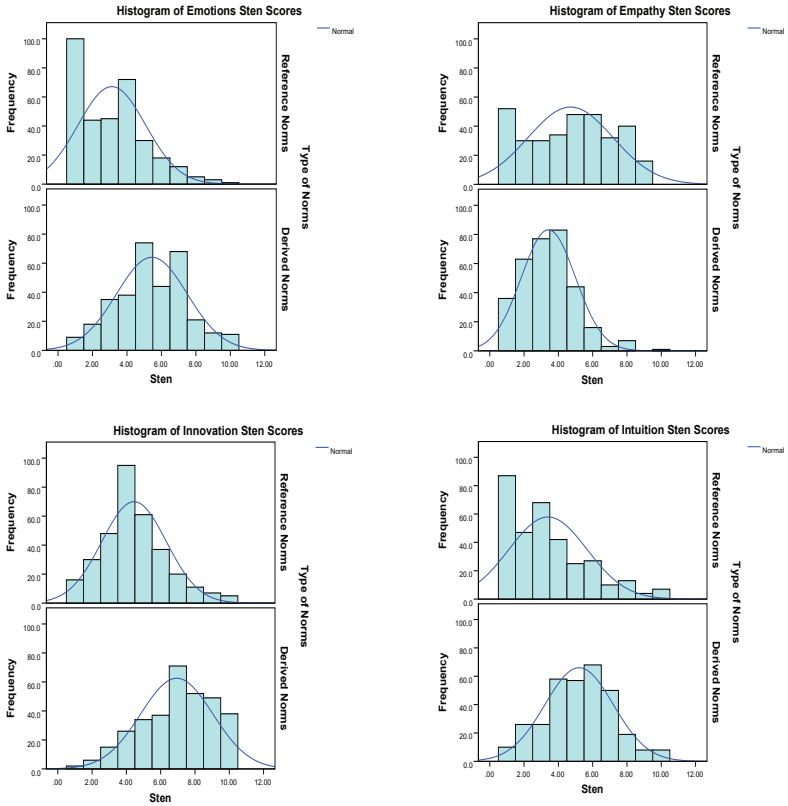
Sten										Age Group
1	2	3	4	5	6	7	8	9	10	
Emotions										
17-19	19-21	21-24	24-26	26-29	29-31	31-34	34-36	36-39	39-42	Teens
19-21	21-23	23-26	26-28	28-31	31-33	33-36	36-39	39-41	41-44	Twenties
20-23	23-25	25-28	28-30	30-32	32-35	35-38	38-40	40-43	43-46	Thirties
21-23	23-25	25-28	28-30	30-33	33-35	35-38	38-41	41-43	43-46	Forties
20-22	22-24	24-27	27-29	29-32	32-34	34-37	37-39	39-42	42-45	Fifties
Empathy										
18-22	22-26	26-29	29-32	32-36	36-38	38-41	41-44	44-46	46-48	Teens
20-24	24-28	28-31	31-35	35-38	38-41	41-43	43-46	46-48	48-50	Twenties
22-26	26-29	29-33	33-36	36-39	39-42	42-45	45-48	48-50	50-52	Thirties
22-26	26-30	30-34	34-37	37-40	40-43	43-46	46-48	48-50	50-52	Forties
22-26	26-30	30-33	33-36	36-40	40-42	42-45	45-48	48-50	50-52	Fifties
Innovation										
16-18	18-19	19-21	21-23	23-25	25-27	27-30	30-32	32-34	34-37	Teens
17-19	19-21	21-23	23-25	25-27	27-29	29-32	32-34	34-37	37-40	Twenties
17-19	19-22	22-24	24-26	26-29	29-31	31-34	34-36	36-39	39-42	Thirties
18-20	20-22	22-24	24-27	27-29	29-31	31-34	34-36	36--39	39-42	Forties
19-21	21-23	23-24	24-26	26-28	28-30	30-32	32-35	35-37	37-40	Fifties
Intuition										
17-19	19-22	22-24	24-27	27-30	30-32	32-35	35-38	38-41	41-45	Teens
19-21	21-24	24-26	26-29	29-31	31-34	34-37	37-40	40-43	43-46	Twenties
21-23	23-25	25-28	28-30	30-33	33-36	36-39	39-42	42-45	45-48	Thirties
22-24	24-26	26-29	29-31	31-34	34-37	37-40	40-43	43-46	46-49	Forties
22-25	25-27	27-29	29-32	32-35	35-37	37-40	40-43	43-46	46-50	Fifties
Motivation										
18-20	20-23	23-25	25-28	28-31	31-34	34-37	37-40	40-43	43-47	Teens
19-22	22-25	25-28	28-31	31-34	34-37	37-40	40-43	43-45	45-48	Twenties
18-23	23-27	27-30	30-34	34-37	37-40	40-42	42-45	45-47	47-49	Thirties
16-22	22-26	26-31	31-35	35-38	38-41	41-43	43-45	45-46	46-47	Forties
12-19	19-24	24-29	29-34	34-37	37-40	40-42	42-43	43-44	44	Fifties
Self-Awareness										
19-21	21-23	23-26	26-29	29-31	31-34	34-38	38-41	41-45	45-48	Teens
19-22	22-25	25-28	28-31	31-34	34-37	37-41	41-44	44-47	47-50	Twenties
19-23	23-27	27-30	30-34	34-37	37-40	40-43	43-46	46-48	48-50	Thirties
17-22	22-27	27-31	31-35	35-38	38-41	41-44	44-46	46-48	48-49	Forties
14-20	20-25	25-30	30-34	34-38	38-41	41-43	43-45	45-46	46-47	Fifties
Social Skills										
15-17	17-20	20-22	22-25	25-27	27-30	30-32	32-35	35-37	37-40	Teens
16-19	19-21	21-24	24-26	26-29	29-31	31-34	34-36	36-39	39-41	Twenties
17-19	19-22	22-24	24-27	27-29	29-32	32-34	34-36	36-39	39-41	Thirties
15-18	18-20	20-23	23-25	25-28	28-30	30-33	33-35	35-38	38-40	Forties
12-14	14-17	17-19	19-22	22-24	24-27	27-29	29-32	32-34	34-37	Fifties

THE_PHILIPPINE_STATISTICIAN_Vol_69_No_2_2020_6x9.indd 29 2/19/2021 10:10:39 AM

Table 7. Kolmogrov-Smirnov Test Results

Factor	Kolmogrov-Smirnov Test Statistic (D)
Emotions	0.488*
Empathy	0.342*
Innovation	0.506*
Intuition	0.424*
Motivation	0.121*
Self-Awareness	0.300*
Social Skills	0.412*

From the histogram of the stens (See Figure 3), it is clear that the reference norms in all factors, excluding Self-Awareness, set high standards for the Filipino sample used in the study. What is considered low in the reference norms is relatively average for the Filipino sample. For Self-Awareness, what is considered average for the reference norms is considered to be approximately low average for the Filipino sample. Relying on the reference norms can lead to an underestimation or an overestimation of the Filipino sample respondents' emotional intelligence abilities.



THE_PHILIPPINE_STATISTICIAN_Vol_69_No_2_2020_6x9.indd 30 2/19/2021 10:10:39 AM

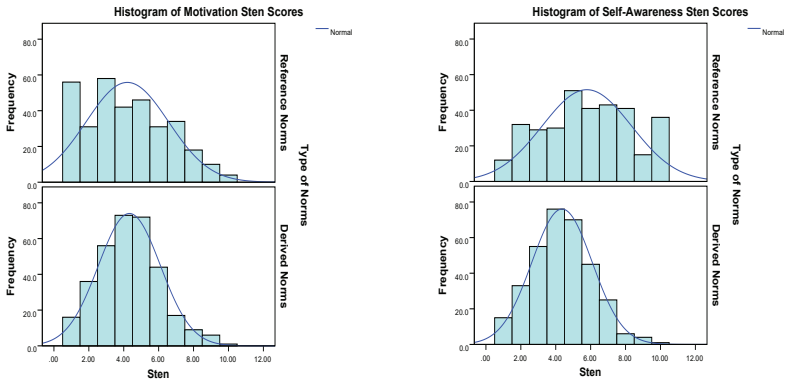


Figure 3. Histogram of Sten Scores

5. General Discussion

Based on Lehnard et al. (2016), normalized standard scores or percentiles can be used as the person location variable; however, given the results, the study showed that the use of stens, rather than percentiles, had produced models with better fit for all seven factors. The study cannot generalize whether other types of normalized standard scores would lead to similar results. In choosing the optimal smoothing parameter, the largest jump in adjusted R-squared and AICs occurred at $k = 2$ for stens for all 7 factors. Although Lehnard et al. (2016) recommended starting at $k = 5$, it was not feasible in the current study not only because the adjusted R-squared and AIC at $k = 5$ no longer differed drastically from smaller smoothing parameters, but also because the predicted values resulting from the $k = 5$ model would likely go beyond the raw score range of the WPQei, considering that variables with higher exponents were involved. Moreover, the adjusted R-squared and AICs around $k = 5$ were already fluctuating. A higher smoothing parameter did not necessarily ensure a better model. Statistically, a higher smoothing parameter can risk overfit; while practically, a higher smoothing parameter can produce predicted raw scores that are greater (or less) than the maximum (or minimum) score of the test.

From the validation analyses, the study showed that the model fit did not drastically change when used in the test set. However, in terms of consistency, only one factor had the same significant variables and coefficient signs in the training and test set. Consistency in terms of significant variables and coefficient signs are important because the Taylor polynomial, and eventually the norm scores, are derived from these two elements. Any difference in the sign and significance can affect the norm score values. From this perspective, the norming method is not robust across different sample sizes. To address this issue, another means of defining the Taylor polynomial was adopted by using the bootstrap estimated coefficients as the constants of the function and the bootstrap confidence interval to identify the significant variables. Bootstrapping is a helpful remedy to these

THE_PHILIPPINE_STATISTICIAN_Vol_69_No_2_2020_6x9.indd 31 2/19/2021 10:10:39 AM

issues given its established benefits in statistical inference. More so, the model defined from the bootstrap results was able to produce raw scores that were within the expected raw score range of the WPQei.

In computing the norm table, the withstanding issue is the repeating values in adjacent stens. Although the iterative method may have solved this problem, specific guidelines for the simpler method are important for practitioners who are not familiar with JavaScript functions or other related software that make use of iterations. Furthermore, the norming technique can produce sten scores that are equivalent to only a single raw score. Typically, each sten is expected to have a range of raw scores. Similar to the previous issue, the iterative method may have addressed the problem; however, considering a higher smoothing parameter may also be another option if feasible. Even so, this appears to be a challenge when using the norming technique.

6. Conclusion and Recommendation

When using the non-parametric continuous norming procedure, the recommended smoothing parameter for WPQei is 2, while the recommended person location variable is the sten instead of percentiles. Future researchers can investigate whether the same conclusion holds true for other types of normalized standard scores. From a practical perspective, choosing the smoothing parameter may pose as a challenge because it has a trial-and-error component, especially since the choice of smoothing parameter varies according to the sample size (Lehnard et al., 2016) as well as the minimum/maximum raw scores of the test. Practitioners should keep in mind that more items are more likely to use higher smoothing parameters compared to tests with lesser items.

Another issue is that the norming technique can produce varying final models across different sample sizes because the significant variables and the coefficient signs (from the stepwise regression) tend to differ accordingly. Bootstrapping was recommended as an alternative means of defining the Taylor polynomial function in order to address the inconsistencies of the norming method and make it more robust. Other difficulties emerged when creating the norm tables. Firstly, the issue of repeating raw scores on adjacent stens was prevalent. Clear guidelines should be established to accommodate practitioners who will not be using the iterative method of computing norm tables. Secondly, the method can unusually produce a sten score that is equivalent to only a single-digit raw score. Using the iterative method or considering a higher smoothing parameter are possible remedies for these issues. Nonetheless, these are possible scenarios that practitioners may have to deal with when using the norming method.

Overall, the non-parametric continuous norming technique is a replicable procedure in practice. Unlike other norming methods, this technique can produce norms from any form of raw score distribution. However, there are some challenges when using the method; such that it is not as robust and smooth-flowing. This study can serve as guide for those who intend to adopt the norming technique in their own practice.

THE_PHILIPPINE_STATISTICIAN_Vol_69_No_2_2020_6x9.indd 32 2/19/2021 10:10:39 AM

References

ANASTASI, A. & URBINA, S. (2009). *Psychological testing* (7th ed.). Singapore: Pearson Education Asia Pte. Ltd.

ANGOFF, W. H. (1984). *Scales, norms, and equivalent score.* New Jersey: Educational Testing Service.

CAMERON, A. (2004). *Work Profile Questionnaire- Emotional Intelligence (WPQei) User's Guide*. Oxford: The Test Agency Limited.

CANTY, A. & RIPLEY, B. (2016). boot: Bootstrap R (S-Plus) Functions. R package version 1.3-18. https://CRAN.R-project.org/package=boot

CHERNICK, M. R. & LABUDDLE, R. A. (2011). *An introduction to bootstrap methods with Applications to R*. New Jersey: John Wiley & Sons, Inc.

CROCKER, L. & ALGINA, J. (1986). *Introduction to classical and modern test theory.* New York: CBS College Publishing.

DAVISON, A. C. & HINKLEY, D. V. (1997). *Bootstrap methods and their application*. New York: Cambridge University Press.

FISCHER R. & MILFONT T. L. (2010). Standardization in psychological research. *International Journal of Psychological Research*, *3*(1), 88–96.

GROTH-MARNAT, G. (2003). *The handbook of psychological assessment* (4th ed.). New Jersey: John Wiley & Sons., Inc.

KAFETSIOS, K. (2004). Attachment and emotional intelligence abilities across the life course. *Personality and Individual Differences*, *37*(1), 129-145. doi: https://doi.org/10.1016/j.paid.2003.08.006

KOMSTA, L. & NOVOMESTKY, F. (2015). moments: Moments, cumulants, skewness, kurtosis and related tests. R package version 0.14. https://CRAN.R-project.org/package=moments

LENHARD, A., LEHNARD, W., SEBASTIAN, S., & SEGERER, R. (2016). A continuous solution to the norming problem. *Assessment*. doi: 10.1177/1073191116656437

MAYER, J. D., CARUSO, D. R., & SALOVEY, P. (1999). Emotional intelligence meets traditional standards for an intelligence. *Intelligence*, *27*, 267-298. http://dx.doi.org/10.1016/S0160-2896(99)00016-1

MOOD, A. M. (1913). *An introduction to the theory of statistics*. New York: McGraw Hill, Inc.

OOSTERHUIS, H., VAN DER ARK, L., & SIJTSMA, K. (2016). Sample size requirements for traditional and regression-based norms. *Assesement*, *23*(2), 191-202. doi: 10.1177/1073191115580638

R CORE TEAM (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

THE_PHILIPPINE_STATISTICIAN_Vol_69_No_2_2020_6x9.indd 33 2/19/2021 10:10:39 AM

REYNOLDS, C. R. & RAMSAY, M. C. (2003). Bias in psychological assessment: An empirical review and recommendations. In J. R. Graham and J. A. Naglieri (Eds.), *Handbook of Psychology Volume 10* (pp.67-94). New Jersey: John Wiley & Sons, Inc.

VAN BREUKELEN, G. J. P., & VLAEYEN, J. W. S. (2005). Norming clinical questionnaires with multiple regression: the pain cognitions list. *Psychological Assessment*, *17*(3), 336–344. doi: 10.1037/1040-3590.17.3.336

ZACHARY, R. A. & GORSUCH, R. L. (1985). Continuous norming: implications for the WAIS-R. *Journal of Clinical Psychology*, *41*, 86-94. 10.1002/1097-4679(198501)41:13.0.CO;2-W

THE_PHILIPPINE_STATISTICIAN_Vol_69_No_2_2020_6x9.indd 34 2/19/2021 10:10:39 AM