Twin Research and Human Genetics (2020), **23**, 247–255
doi:10.1017/thg.2020.64

CAMBRIDGE UNIVERSITY PRESS

Article

Psychometric and Classification Properties of the Peas in a Pod Questionnaire

Ally R. Avery[1], Eric Turkheimer[2], Siny Tsang[1] and Glen E. Duncan[1]

[1]Department of Nutrition and Exercise Physiology, Washington State University, Spokane, WA, USA and [2]Department of Psychology, University of Virginia, Charlottesville, VA, USA

Abstract

We examined the item properties of the Two Peas Questionnaire (TPQ) among a sample of same-sex twin pairs from the Washington State Twin Registry. With the exception of the 'two peas' item, three of the mistakenness items showed differential item functioning. Results showed that the monozygotic (MZ) and dizygotic (DZ) pairs may differ in their responses on these items, even among those with similar latent traits of similarity and confusability. Upon comparing three classification methods to determine the zygosity of same-sex twins, the overall classification accuracy rate was over 90% using the unit-weighted pair zygosity sum score, providing an efficient and sufficiently accurate zygosity classification. Given the inherent nature of twin-pair similarity, the TPQ is more accurate in the identification of MZ than DZ pairs. We conclude that the TPQ is a generally accurate, but by no means infallible, method of determining zygosity in twins who have not been genotyped.

Keywords: Zygosity; item factor analysis; differential item functioning; classification; latent class analysis

(Received 2 July 2020; accepted 3 July 2020; First Published online 10 August 2020)

The earliest twin studies (e.g. Merriman, 1924), conducted before World War II, were based on small samples that were studied in person by the investigator. Zygosity could be determined based on either clinical impression or blood groups. It was only when large twin registries were established in Scandinavia that it became necessary to diagnose zygosity remotely using self-report questionnaires. The first systematic approach to the problem was undertaken in the Swedish Twin Register (STR; Magnusson et al., 2013), who asked participants whether they were as 'lika som bär' (alike as berries). This is why the logo of the STR is a pair of cherries ('korsbar' in Swedish).

English versions of the Swedish questionnaire translated the expression as 'alike as two peas in a pod', and the peas in a pod question have in the years since become the centerpiece of zygosity questionnaires, which are often known as 'peas in a pod questionnaires'. Although no universal standard for such questionnaires has ever emerged, the item about peas is usually combined with a series of questions about whether the twins are confused by parents, family members and acquaintances. Many studies have demonstrated that self-report questionnaires of this kind can make accurate decisions about zygosity when validated against blood markers or genotyping (accuracy rate ranges from 92.4% to 98.8%; e.g. Eisen et al., 1989; Forsberg et al., 2010; Jackson et al., 2001; Jarrar et al., 2018; Magnus et al., 1983; Magnusson et al., 2013; Ohm Kyvik & Derom, 2006; Reed et al., 2005; Song et al., 2010). There is no universal standard for these items, and twin researchers and/or registries have used various forms of these collection of items to assess zygosity. In this article, we refer to our particular version of the questionnaire as the Two Peas Questionnaire (TPQ).

It is somewhat surprising that no systematic examination of the psychometrics of the TPQ has ever been conducted. In fact, the questionnaire is more than just a simple list of questions that can be used with a cutoff to diagnose zygosity; it is a psychological measurement instrument, designed to measure self-reported subjective impressions of similarity and confusability. The validity of the questionnaire as a tool for classification is closely tied to its measurement properties.

There are several reasons to expect that the psychometrics of the TPQ and its application to classification would be less than perfectly straightforward. First, the questionnaire is by design administered to disparate groups of individuals, that is, monozygotic (MZ) and dizygotic (DZ) twins, who might be expected to have different reactions to questions about their similarity and confusability. Second, there is an asymmetry in the way biological differences reflect on zygosity; even small differences are sufficient to demonstrate that a pair of twins is DZ, whereas a high degree of similarity is not sufficient to demonstrate that a pair is MZ. For example, twin pairs with different eye colors are almost certainly DZ, but pairs with the same eye color are not certain to be MZ. This asymmetry leads to an expectation of a difference in the distribution of responses to the TPQ in MZ and DZ twins. When the questionnaire is used as a classification instrument, it will usually be the case that prior probabilities favor a pair being MZ. Identical twins are often easier to ascertain within twin samples, but even if this is not the case in a particular sample, opposite-sex twins will be DZ twins and can be classified without the use of the questionnaire.

Author for correspondence: Ally R. Avery, Email: ally.avery@wsu.edu

Cite this article: Avery AR, Turkheimer E, Tsang S, and Duncan GE. (2020) Psychometric and Classification Properties of the Peas in a Pod Questionnaire. *Twin Research and Human Genetics* **23**: 247–255, https://doi.org/10.1017/thg.2020.64

Finally, there is reason to expect that responses to the questionnaire will vary according to age. Both classic (Scarr & McCartney, 1983) and more recent (Beam & Turkheimer, 2013) analyses show that twins become more different as they age, and that DZ pairs do so more rapidly than MZ pairs.

We report a series of psychometric and classificatory analyses in a large sample of twins who have been administered a TPQ, and a smaller subsample who have been genotyped to provide a biological criterion for zygosity. We estimate item factor analysis (IFA) parameters for the psychometric properties of the questionnaire in the MZ and DZ groups and use them to identify differential item functioning (DIF) across groups. We then estimate the distributions of the latent similarity parameters in the two groups and explore several classification models based on the IFA model and methods based on latent class analysis (LCA).

Study 1

The primary goal of study 1 was to examine the item parameters of the TPQ among a sample of same-sex adult twin pairs with DNA-based zygosity. We used IFA models to examine potential DIF in the TPQ items between MZ and DZ twin pairs. IFA models describe the association between the latent trait level (i.e. underlying trait of being identical) and item scores (i.e. scores on the TPQ), allowing DIF analyses that are not affected by potential differences in the latent trait distributions across groups (Embretson & Reise, 2000).

Methods

Participants

The current study utilized data from 753 same-sex adult twin pairs (33.9% men, 66.1% women) enrolled in the Washington State Twin Registry (WSTR) with DNA-based zygosity (72.4% MZ, 27.6% DZ). The WSTR is a community-based registry of twin pairs primarily recruited through Washington State Department of Licensing records. Details regarding the recruitment procedures of the WSTR and additional information are reported elsewhere (Duncan et al., 2019). Participants in this study were recruited into the WSTR between 2002 and 2014.

DNA Determination of Zygosity

DNA was extracted from twins using either whole blood or saliva (buccal cells). Zygosity was determined by using either the AmpFlSTR® Identifiler® Plus PCR Amplification Kit or the PowerPlex® 16 HS System, per manufacturer's instructions. The two methods are nearly identical (Hannelius et al., 2007; Yang et al., 2006). These kits are short tandem repeat multiplex assays that amplify 15 tetranucleotide repeat loci and the amelogenin sex-determining marker in a single PCR amplification. Thirteen of the required loci (CSF1PO, FGA, TH01, TPOX, vWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51 and D21S11) for the Combined DNA Index System are included (Budowle et al., 1999). Two additional loci, D2S1338 and D19S433, are included. The combination of these 15 loci along with the amelogenin marker is consistent with zygosity tests conducted elsewhere (Yang et al., 2006). When comparing the twins with one another, DZ twins match on 25%–75% of the sites, whereas MZ twins match on 100% of the sites. Zygosity determination for twin pairs in this study was performed between 2009 and 2017.

Two Peas Questionnaire

Five items about childhood similarity were included in the WSTR enrollment survey. The 'two-peas' item, 'When you were children, were you and your twin as alike as two peas in a pod or of ordinary family resemblance?', has been used by twin registries for many years and is a reliable predictor of zygosity (Eisen et al., 1989; Magnus et al., 1983; Reed et al., 2005; Sarna et al., 1978). Four mistakenness items ask, 'When you were children, did the following people (parents, other relatives, teachers, and strangers) have difficulty telling you and your twin apart?' (Buchwald et al., 1999; Eisen et al., 1989; Magnus et al., 1983; Reed et al., 2005). There are four response categories for each of the mistakenness items (1 = *never confused*, 2 = *rarely confused*, 3 = *sometimes confused*, 4 = *always confused*). For ease of interpretation, these four mistakenness items are subsequently referred to as 'parents', 'relatives', 'teachers' and 'strangers', respectively.

Statistical Analysis

We used IFA to estimate the item parameters of the 10 items (i.e. 5 items from each twin, 10 items per twin pair) in the TPQ. The 10 items were operationalized as indicators of the underlying latent trait (θ) of being similar and easily confused (i.e. more MZ-like), with higher levels reflecting stronger endorsement of being identical, whereas lower levels reflecting endorsement of being less identical. Considering that the items in the TPQ consist of ordinal response options, IFA is an alternative to the common linear factor model when item responses are categorical in nature (Wirth & Edwards, 2007). One factor-loading parameter was estimated for each of the five items (λ_1 – λ_5). One threshold parameter (τ_1) was estimated for the dichotomous 'two peas' item, and three threshold parameters (τ_{21}, τ_{22}, τ_{23}, ... τ_{51}, τ_{52} and τ_{53}) were estimated for each of the remaining four items, each with four response categories. All factor loadings and threshold parameters were constrained to be the same within twin pairs, and item covariances within twin pairs were allowed to differ between MZ and DZ twin pairs. Participants were designated as MZ and DZ using DNA-based zygosity.

First, we fit a 'free-baseline' model in which the factor loadings of a reference item (our selection of the reference item is described below) were fixed to 1, and the threshold parameters were constrained to be equal between MZ and DZ pairs (Stark et al., 2006). The factor loadings and threshold parameters for the remaining four items were allowed to differ between MZ and DZ pairs. In order to detect items with DIF, we fit four constrained models where, in addition to the reference item, factor loadings and threshold parameters of each item, one at a time, were simultaneously constrained to be equal between MZ and DZ twins. Items with DIF were identified by comparing the changes in chi-square statistics. To control for type I errors due to multiple comparisons, a Bonferroni-corrected critical p value (.05/4 = .0125) was used.

To identify the reference item(s), we fit a fully constrained model in which the factor loadings and threshold parameters of all items were constrained to be equal between MZ and DZ pairs. Next, we fit a series of augmented models by freeing the factor loadings and threshold parameters one item at a time. The item(s) that did not result in a statistically significant increase in model fit when the parameters were allowed to differ between MZ and DZ twins was identified as the reference item(s) (Stark et al., 2006). To control for type I errors due to multiple comparisons, a Bonferroni-corrected critical p value of (.05/5 = .01) was used.

https://doi.org/10.1017/thg.2020.64 Published online by Cambridge University Press

Table 1. Selected demographic characteristics of the Washington State Twin Registry (WSTR) twin pairs included in this study

	Twins with DNA-based zygosity	MZ	DZ	Twins without DNA-based zygosity
Number of twin pairs	753	545 (72.4%)	208 (27.6%)	6368
Gender (% men)	33.9	29.8	26.5	35.9
Race (% White)	86.7	88.1	83.2	90.3
Age at enrollment	$M = 29.7$, $SD = 14.7$, range 18.0–85.5	$M = 30.1$, $SD = 14.7$, range 18.0–85.5	$M = 28.5$, $SD = 14.8$, range 18.1–70.5	$M = 40.2$, $SD = 18.2$, range 18.0–92.4

MZ, monozygotic twins; DZ, dizygotic twins.

Model fit indices reported include the comparative fit index (CFI), Tucker-Lewis index (TLI), root mean square error of approximation (RMSEA) and standardized root mean squared residual. Descriptive statistics were performed using R version 3.5.3 (R Development Core Team, 2015), and IFA models were performed using Mplus version 8.1 (Muthén & Muthén, 2012).

Results

Descriptive Statistics

Of the 753 pairs of same-sex twins in this study, there were 545 (72.4%) MZ and 208 (27.6%) DZ twin pairs as determined by genotyping. Selected demographic characteristics of twin pairs in this study are presented in Table 1.

Descriptive statistics of the five TPQ items are shown in Table 2. For the 'two peas' item, most of the MZ twins (93%) reported that they were 'as alike as two peas in a pod', whereas the majority of the DZ twins (84%) responded that they were 'of ordinary family resemblance' when they were children. Concordance rates of the 'two peas' item are presented in Supplementary Table 1. For the four mistakenness items, larger proportions of MZ twins reported being confused by teachers and strangers (68% and 91% always confused, respectively) than by parents and other relatives (12% and 49% always confused, respectively) when they were children. On the other hand, small proportions of DZ twins reported being confused by teachers and strangers (11% and 20% always confused, respectively), and even smaller proportions reported being confused by parents and other relatives (3% and 6% always confused, respectively).

Differential Item Functioning

Identify reference item. In order to identify the reference item, we fit a fully constrained model in which the factor loadings and threshold parameters of all items were constrained to be equal between MZ and DZ twins. The model was of acceptable fit (CFI = .980, TLI = .975, RMSEA = .067, 90% CI = .055, .078, SRMS = .060). Next, we fit a series of augmented models in which, one item at a time, the factor loadings and threshold parameters were simultaneously allowed to differ between MZ and DZ twins. Chi-square tests showed that there was no statistically significant improvement in model fit when the parameters for the 'peas' or 'strangers' item were allowed to differ between MZ and DZ twin pairs (Supplementary Table 2). Considering that the change in model fit was the smallest when the parameters for the 'strangers' item differed between MZ and DZ twins, the 'strangers' item was used as the reference item in the subsequent analyses.

Table 2. Descriptive statistics of the Two Peas Questionnaire items (individual twin's responses)

	Zygosity	Ordinary resemblance (%)	Two peas in a pod (%)		
Two peas	MZ	7	93	–	–
	DZ	84	16	–	–
	Unknown	33	67	–	–
		Never confused (%)	Rarely confused (%)	Sometimes confused (%)	Always confused (%)
Parents	MZ	24	34	30	12
	DZ	83	11	3	3
	Unknown	45	25	22	8
Relatives	MZ	3	10	38	49
	DZ	57	24	13	6
	Unknown	24	13	30	34
Teachers	MZ	1	4	27	68
	DZ	49	21	19	11
	Unknown	21	8	24	47
Strangers	MZ	1	2	7	91
	DZ	41	17	22	20
	Unknown	19	6	10	65

MZ, monozygotic twins; DZ, dizygotic twins; unknown, twin pairs with no DNA-based zygosity. Note: Two peas: When you were children, were you and your twin as alike as two peas in a pod or of ordinary family resemblance? Parents: When you were children, how often did your parents had difficulty telling you apart? Relatives: When you were children, how often did other relatives had difficulty telling you apart? Teachers: When you were children, how often did teachers had difficulty telling you apart? Strangers: When you were children, how often did strangers had difficulty telling you apart?

Test for DIF. To test for DIF among self-report zygosity items, we first fit a 'free baseline' model where the factor loadings of the 'strangers' item (the reference item identified above) were fixed to 1, and the threshold parameters were constrained to be equal between MZ and DZ. The factor loadings and threshold parameters for the remaining four items — 'two peas', 'parents', 'relatives' and 'teachers' — were allowed to differ between MZ and DZ pairs. As shown in Table 3, the model fit was good (CFI = .990, TLI = .985, RMSEA = .052, 90% CI = .038, .066, SRMS = .054) and was a better fit than the fully constrained model, $\chi^2(14) = 70.107$, $p < .001$.

Next, we fit four constrained models in which, one item at a time, in addition to the 'strangers' item, factor loadings and threshold parameters of each item were simultaneously constrained to be equal between MZ and DZ pairs. The model fit of these constrained models was compared against the 'free-baseline' model using chi-square tests (Supplementary Table 2). With the exception of the 'two peas' item, there was a statistically significant decrease in model fit when the item parameters were constrained to be equal between MZ and DZ pairs, suggesting DIF between MZ and DZ twins in the 'parents', 'relatives' and 'teachers' items.

https://doi.org/10.1017/thg.2020.64 Published online by Cambridge University Press

Table 3. Estimated factor loadings and thresholds of the free-baseline model for the self-report zygosity items

		MZ		DZ	
		Est	*SE*	Est	*SE*
Loadings					
Two peas	λ_1	1.146	.12	.911	.12
Parents	λ_2	.741	.08	.814	.12
Relatives	λ_3	1.052	.10	.848	.09
Teachers	λ_4	1.038	.09	1.004	.08
Strangers	λ_5	1	–	1	–
Thresholds					
Two peas	τ_{11}	−1.499	.07	−.999	.29
Parents	τ_{21}	−.693	.05	−.827	.29
	τ_{22}	.217	.05	−.192	.31
	τ_{23}	1.197	.06	.142	.33
Relatives	τ_{31}	−1.915	.09	−1.682	.21
	τ_{32}	−1.115	.06	−.970	.22
	τ_{33}	.034	.05	−.273	.24
Teachers	τ_{41}	−2.167	.10	−2.232	.19
	τ_{42}	−1.591	.07	−1.672	.19
	τ_{43}	−.461	.05	−.957	.19
Strangers	τ_{51}	−2.430	.08	−2.430	.08
	τ_{52}	−1.991	.08	−1.991	.08
	τ_{53}	−1.345	.06	−1.345	.06
Mean		0	–	−2.189	.10
Variance		.429	.07	.520	.07
Model fit					
RMSEA (90% CI)		.052 (.038, .066)			
CFI		.990			
TLI		.985			
SRMR		.054			

MZ, monozygotic twins; DZ, dizygotic twins; *SE*, standard error; RMSEA, root mean square error approximation; CFI, comparative fit Index; TLI, Tucker-Lewis index; SRMR, standardized root mean square residual. Note: Only parameters of one twin are shown here, as all item parameters are constrained to be the same within twin pairs. 'Strangers' is used as the referent item, with the factor loadings fixed to 1 and threshold parameters constrained to be equal between MZ and DZ twins.

We illustrate the similar item functioning (i.e. no DIF) of 'two peas' for MZ and DZ twins using category response curves (CRCs). As shown in Figure 1, the probabilities that MZ and DZ twins responded they were 'two peas in a pod' or 'of ordinary resemblance' were similar. For example, at $\theta = 0$ (i.e. the average latent trait level of similarity and confusability), there was a 98.8% chance that MZ twins responded they were 'two peas in a pod', but only 1.2% chance that they identified themselves as 'of ordinary resemblance'. At the same latent trait level ($\theta = 0$), DZ twins were also more likely to respond that they were 'two peas in a pod' (90.7%) and less likely to identify themselves as of 'ordinary resemblance' (9.3%).

DIFs of the other three items are illustrated using CRCs (Supplementary Figure 1). Among twins with similar levels of the latent trait of being identical, DZ twins were more likely to respond that other people had difficulty telling them apart than MZ twins. For instance, at $\theta = 0$, MZ twins were likely to respond that they were 'rarely confused' (38.4%) and 'sometimes confused' (31.6%) by parents, whereas DZ twins were more likely to respond that they were 'always confused' (43.1%) by parents. Likewise, MZ twins at $\theta = 0$ were more likely to respond that they are 'always confused' (48.2%) or 'sometimes confused' (45.6%) by relatives, whereas DZ twins at $\theta = 0$ were most likely to respond that they are 'always confused' by relatives.

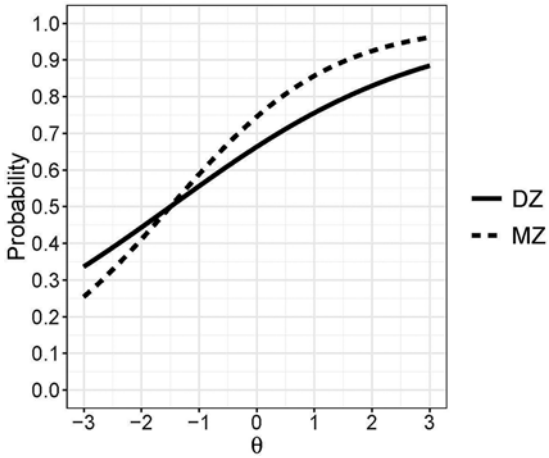

Fig. 1. Category response curves (CRCs) of the 'two peas' item by zygosity among twin pairs with DNA-based zygosity.

Discussion

In study 1, we estimated the item parameters for the TPQ items using IFA models and examined whether there was DIF between MZ and DZ twin pairs. Results showed no loss of model fit when the 'two peas' item parameters were constrained to be equal across zygosity, suggesting the 'two peas' item functions similarly for MZ and DZ twins. Our analyses showed DIF in three of the mistakenness items on the TPQ, 'parents', 'relatives' and 'teachers'. For these items, the probabilities of responses may differ not only by individuals' underlying trait of being similar and confusable (i.e. more MZ-like or more DZ-like) but also by their actual zygosity (i.e. true MZ or true DZ twins, based on genotyping).

When twin pairs' responses are used to classify twins with unknown zygosity into MZ or DZ pairs, it is possible that DIF in TPQ items may affect which twin pairs are assigned as MZ or DZ twin pairs. We followed up the current findings with a second study in which we explored several classification methods for zygosity assignment to establish an effective method to determine zygosity assignments among twin pairs that have not yet been genotyped.

Study 2

In study 2, we aimed to investigate three classification methods used to assign twins into MZ and DZ pairs, based on their responses on the TPQ. Zygosity of twin pairs was classified based on their unit-weighted pair zygosity sum (PZS) score, item response probabilities from an IFA model and item response probabilities from a LCA model.

https://doi.org/10.1017/thg.2020.64 Published online by Cambridge University Press

Methods

Participants

Twin pairs included in this study were the 753 twin pairs with DNA-based zygosity described in study 1, as well as 6368 same-sex adult twin pairs (35.9% men, 64.1% women) enrolled in the WSTR without DNA-based zygosity (Table 1). The recruitment procedures of these twin pairs were like those described in study 1.

Two Peas Questionnaire

The TPQ described in study 1 was also used in study 2.

Statistical Analysis

Unit-weighted PZS scores. Using the twins' responses on the TPQ, we created a unit-weighted PZS score for each twin pair. The four mistakenness questions were first rescaled to the same scale as the dichotomous two peas item (0 = 0; 1 = .33, 2 = .67, 3 = 1). The PZS scores were computed by summing the scores of the 10 items (i.e. 5 items per twin) in the TPQ. PZS score ranged from 0 to 10, with higher scores reflecting higher degrees of similarity and confusability. For twin pairs with missing items, PZS scores were rescaled by:

$$\text{PZS} = \frac{\text{PZS}}{\text{Total number of non–missing items}} \times 10.$$

Probabilities of zygosity (MZ/DZ) from PZS scores. We fit a logistic regression model to estimate the probabilities of zygosity (MZ/DZ) using the PZS scores among twin pairs with DNA-based zygosity. To determine the optimum PZS cutoff value to classify twin pairs into MZ and DZ twin pairs, we performed cross-validation using 75% of the data randomly sampled as the training set, and the remaining 25% of the data used as the testing set. The optimum cutoff was the PZS value with the maximum overall classification accuracy rate (i.e. real MZ/DZ pairs correctly classified as MZ/DZ pairs). This procedure was repeated 1000 times. The final PZS cutoff value was determined by taking the average of the PZS cutoffs from the 1000 cross-validations. Subsequently, twin pairs were assigned as MZ and DZ twin pairs using the final PZS cutoff value; this zygosity assignment was referred to as the 'PZS zygosity'.

IFA model for MZ and DZ twins. We used IFA to estimate the item parameters of the 10 items (i.e. 5 items per twin) in the TPQ, separately for MZ and DZ twin pairs with DNA-based zygosity. The 10 items were operationalized as indicators of the underlying latent trait (θ) of similarity and confusability (i.e. more identical or MZ-like), with higher levels reflecting stronger endorsement of similarity and confusability, whereas lower levels reflecting weaker endorsement of the latent trait. One factor loading parameter was estimated for each of the five items (λ_1 – λ_5). One threshold parameter (τ_1) was estimated for the dichotomous 'two peas' item, and three threshold parameters (τ_{21}, τ_{22}, τ_{23}, ... τ_{51}, τ_{52} and τ_{53}) were estimated for each of the remaining four items, each with four response categories. To estimate all factor loadings and threshold parameters, the mean and variance of the latent zygosity factor was fixed to 0 and 1, respectively. All factor loadings and threshold parameters were constrained to be the same for corresponding items within twin pairs, and residual item covariances within twin pairs were estimated. The IFA model was fit separately for MZ and DZ twin pairs.

Probabilities of zygosity (MZ/DZ) from IFA model. Using the estimated item parameters from the IFA models, the probabilities of each response category for each item were computed across the latent trait distribution using the Gaussian quadrature procedure (Embretson & Reise, 2000). We computed the probability of getting a particular response vector ($\underline{X}_p$) in a random sample by integrating the IFA model estimation over the range of latent trait distribution:

$$P\left(\underline{X}_p\right) = \int \prod_i P^{x_{ip}}\ Q^{1-x_{ip}} g(\theta) d\theta$$

where $g\theta$ is the probability density of the latent trait θ.

As the item parameters were estimated separately for MZ and DZ twin pairs, two sets of probabilities were computed, one for MZ and one for DZ twins. We computed the response pattern likelihoods for each twin pair based on their responses on the TPQ. The probability of a particular response pattern was obtained by multiplying the response likelihoods for each of the 10 items. For ease of computation, likelihoods were log-transformed into log-likelihoods. As such, the log-likelihood of a particular response pattern was the sum of the log-likelihoods of the 10 items.

For each twin pair, we obtained the log-likelihoods of the pair being MZ (lnL_{MZ}) or DZ (lnL_{DZ}) twins. The log-likelihoods are monotonic transformations of the probabilities of the pair being an MZ or DZ pair. By taking the difference between the two log-likelihoods ($\Delta lnL = lnL_{MZ} - lnL_{DZ}$), twin pairs with larger ΔlnL had higher probabilities of being MZ (i.e. $lnL_{MZ} > lnL_{DZ}$), suggesting they had higher levels of similarity and confusability (i.e. more likely to be MZ twins). Those with smaller ΔlnL (i.e. $lnL_{DZ} > lnL_{MZ}$) had higher probabilities of being DZ twins. To determine the optimum cutoff value for zygosity classification, we computed the overall classification accuracy rate at each ΔlnL. The optimum cutoff value was determined at the ΔlnL with the maximum accuracy rate; if the maximum accuracy rate occurred at multiple ΔlnL, we took the average of all ΔlnLs as the final cutoff value. Twin pairs were assigned as MZ or DZ twins using the final cutoff value, and we referred to this zygosity assignment as the 'IFA zygosity'.

Latent class analysis. LCA (McCutcheon, 1987) is a type of mixture modeling technique that aims to describe the heterogeneity in a population by identifying substantively meaningful subgroups. These otherwise unobserved subgroups, or latent classes, are characterized by similar patterns of responses on measured categorical indicators (Collins & Lanza, 2010). Two sets of parameters are estimated from LCA, the latent class membership probabilities and the item response probabilities. The latent class membership probabilities represent the likelihood a participant or a response pattern belongs to the latent class. The probabilities of these latent class memberships sum to 1, within rounding error. The item response probabilities refer to the likelihood of each response category to each item for each latent class. We used LCA to estimate the item parameters of the 10 items (i.e. 5 items per twin) in the TPQ among twin pairs without DNA-based zygosity.

Probabilities of zygosity (MZ/DZ) from LCA model. Using the estimated item response probabilities from the LCA models, the response pattern likelihoods for each twin pair were computed, following similar procedures outlined above for those from the IFA models. The corresponding zygosity assignments were referred to as 'LCA zygosity'.

https://doi.org/10.1017/thg.2020.64 Published online by Cambridge University Press

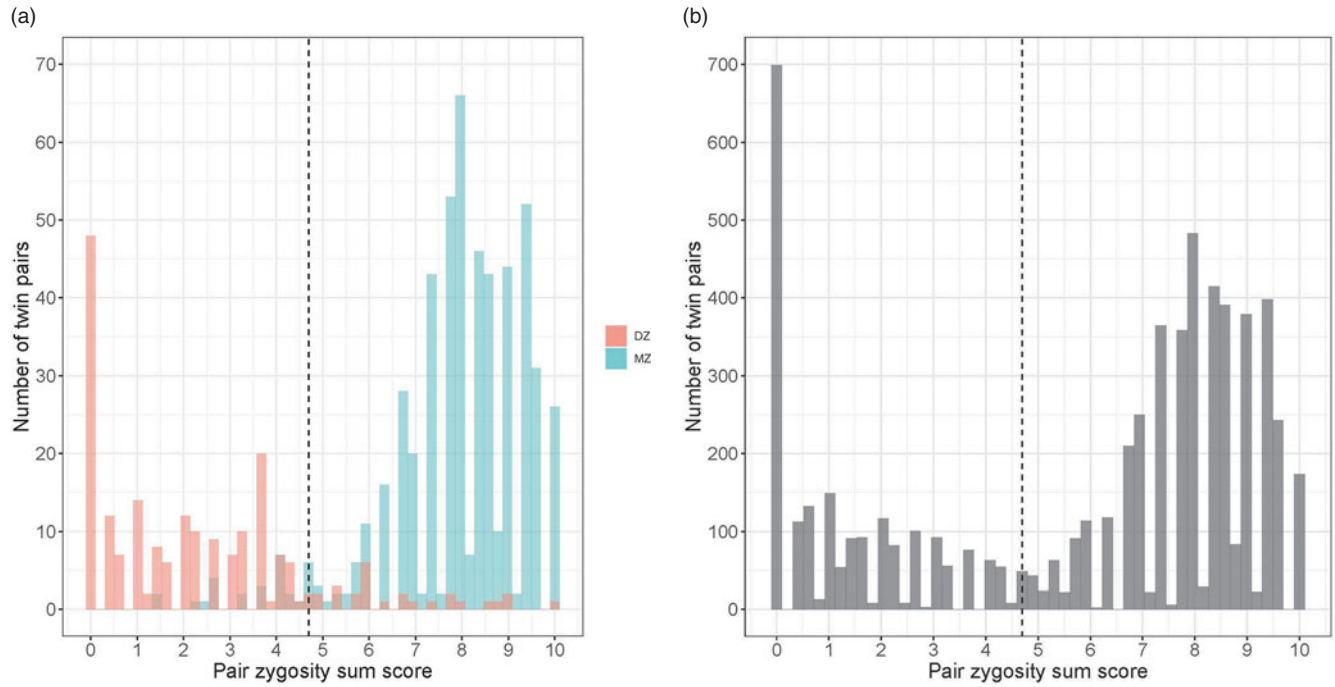

Fig. 2. Distribution of unit-weighted pair zygosity sum (PZS) score among twin pairs with and without DNA-based zygosity. Dashed line indicates the optimum cutoff value at PZS = 4.7. (a) Twin pairs with DNA-based zygosity. (b) Twin pairs without DNA-based zygosity.

Classification accuracy. We evaluated the classification accuracy of the three zygosity assignments — (1) PZS zygosity, (2) IFA zygosity and (3) LCA zygosity — among the twin pairs with DNA-based zygosity. For each zygosity assignment, we computed the classification accuracies for MZ and DZ twin pairs (i.e. the proportion of true MZ/DZ twins correctly classified as MZ/DZ twins). As the item response probabilities for the IFA zygosity assignments were estimated in the same sample (twins with DNA-based zygosity), cross-validation was not possible. To obtain out-of-sample estimates of classification accuracy, the item response probabilities for the LCA zygosity assignments were estimated in the sample without DNA-based zygosity and subsequently validated in the sample with DNA-based zygosity.

Classification consistency. As the true zygosity of twin pairs without DNA zygosity was unknown, we evaluated the extent to which the three zygosity assignments were consistent across one another. We computed the proportion of twin pairs that was consistently assigned as MZ or DZ twin pairs, as well as the proportion of twin pairs that did not have consistent zygosity assignment across the three methods. Reliability across zygosity assignment was evaluated using Fleiss' kappa (Fleiss, 1971).

Results

Descriptive Statistics

Selected demographic characteristics and descriptive statistics of the five self-report zygosity items are shown in Tables 1 and 2, respectively.

Zygosity assignments from PZS scores. Among twin pairs with DNA-based zygosity, the average PZS scores were substantially higher in MZ pairs, and the variance of PZS scores was higher in DZ pairs: 7.9 ($SD = 1.5$) and 2.4 ($SD = 2.2$) for MZ and DZ pairs, respectively. The average PZS score among twin pairs without DNA-based zygosity was 5.9 ($SD = 3.4$). The distribution of PZS scores among twin pairs with and without DNA-based zygosity is illustrated in Figure 2 in which the scores of the DZ pairs are more variable and more skewed in the direction of similarity.

Using data from twin pairs with DNA-based zygosity, the optimum cutoff PZS value at which the highest classification accuracy rate was obtained for each of the 1000 cross-validated logistic regression models is illustrated in Supplementary Figure 2. The average optimum cutoff value was at PZS = 4.7 ($SE = .03$). PZS zygosity was obtained by assigning twin pairs with PZS $\geq$ 4.7 as MZ twins and those with PZS < 4.7 as DZ twins (Tables 4 and 5).

Zygosity assignments from IFA and LCA models. IFA item response probabilities were obtained from the sample with DNA-based zygosity (Supplementary Table 5). For each twin pair, the difference between the response pattern likelihoods of being MZ and DZ was computed. The overall maximum classification accuracy rate (93.5%) was obtained at $\Delta lnL_{IFA} = -4.5$ and -4.4; thus, we took the average of these values and determined the optimum cutoff value at $\Delta lnL_{IFA} = -4.45$ (Supplementary Figure 3). The distribution of ΔlnL_{IFA} obtained from the IFA model is illustrated in Figure 3. Descriptive statistics of the zygosity assignment based on IFA item response probabilities ('IFA zygosity') are presented in Tables 4 and 5.

Similarly, LCA item response probabilities were obtained from the sample without DNA-based zygosity (Supplementary Table 6). The difference between the response pattern likelihoods of being MZ and DZ was computed. The overall maximum classification accuracy rate (93.8%) was obtained at $\Delta lnL_{LCA} = 1.0$ and 1.4; thus, we took the average of these values and determined the optimum cutoff value at $\Delta lnL_{LCA} = 1.2$ (Supplementary Figure 4). The distribution of ΔlnL_{LCA} obtained from the LCA model is illustrated in Figure 4. Descriptive statistics of the zygosity assignment based on LCA item response probabilities ('LCA zygosity') are presented in Tables 4 and 5.

https://doi.org/10.1017/thg.2020.64 Published online by Cambridge University Press

Table 4. Comparison of DNA-based zygosity with three zygosity classifications among twin pairs with DNA-based zygosity

PZS zygosity				
DNA-based zygosity	MZ	DZ	Accuracy	Overall accuracy
MZ	516	29	94.7%	92.7%
DZ	26	182	87.5%	
IFA zygosity				
DNA-based zygosity	MZ	DZ	Accuracy	Overall accuracy
MZ	521	24	95.6%	93.5%
DZ	25	183	88.0%	
LCA zygosity				
DNA-based zygosity	MZ	DZ	Accuracy	Overall accuracy
MZ	517	28	94.9%	93.6%
DZ	20	188	90.4%	
Consistent across classification methods				
DNA-based zygosity	MZ	DZ	Inconsistent	
MZ	512	23	10	
DZ	16	178	14	

MZ, monozygotic twins; DZ, dizygotic twins; PZS zygosity, zygosity assignment based on twin pairs' pair zygosity sum (PZS) scores; IFA zygosity, zygosity assignment based on item factor analysis (IFA) model; LCA zygosity, zygosity assignment based on latent class analysis (LCA) model.

Table 5. Comparison of three zygosity classifications among twin pairs without DNA-based zygosity

	MZ		DZ	
	n	%	*n*	%
PZS zygosity	4306	67.6	2061	32.4
IFA zygosity	4355	68.4	2014	31.6
LCA zygosity	4254	66.8	2115	33.2
Consistent across classification methods[a]	4212	66.2	1991	32.3

MZ, monozygotic twins; DZ, dizygotic twins; PZS zygosity, zygosity assignment based on twin pairs' pair zygosity sum (PZS) scores; IFA zygosity, zygosity assignment based on item factor analysis (IFA) model; LCA zygosity, zygosity assignment based on latent class analysis (LCA) model.
Note: [a]The percentages did not add up to 100% as 164 (2.6%) twin pairs were not consistently classified as MZ or DZ twins.

Classification accuracy. We compared the three zygosity assignments against the DNA-based zygosity among the twin pairs with DNA-based zygosity. The overall accuracy ranged from 92.7% (PZS zygosity) to 93.6% (LCA zygosity). Among MZ twins, 94.7% (PZS zygosity) to 95.6% (IFA zygosity) were correctly assigned as MZ. The classification accuracy was lower among DZ twins, with 87.5% (PZS zygosity) to 90.4% (LCA zygosity) correctly assigned as DZ. Fleiss' kappa = .947 indicated excellent consistency across the three zygosity assignments; 512 (93.9%) MZ pairs were consistently correctly classified as MZ, and 178 (85.6%) DZ pairs were consistently correctly classified as DZ.

Classification consistency. Among the twin pairs without DNA-based zygosity, 66.8% (LCA zygosity) to 68.4% (IFA zygosity) were classified as MZ twins, and 31.6% (IFA zygosity) to 33.2% (LCA zygosity) were classified as DZ twins. The three zygosity assignments were highly consistent, with 6203 (98.5%) twin pairs consistently assigned as MZ (4212 pairs; 66.2%) and DZ (1991 pairs; 32.3%) twins, respectively. Fleiss' kappa = .961 indicated excellent consistency across the three zygosity assignments.

Discussion

In study 2, we examined zygosity assignments predicated on three classification methods. Among twin pairs with DNA-based zygosity, classification accuracies were consistently high. Zygosity assignments were highly consistent among twin pairs with and without DNA-based zygosity. Although the accuracies of zygosity assignment were improved when using more sophisticated classification methods (i.e. IFA and LCA), the difference was minimal (<1%; <10 twin pairs) as compared to zygosity assignment from a simple logistic regression model.

We noted that among the 753 twin pairs with DNA-based zygosity, 39 pairs (23 MZ, 16 DZ) were consistently misclassified by all three methods. As illustrated in Supplementary Figure 5, the response patterns of these twin pairs were not consistent with those of their respective zygosity. Our classification methods depended on participants' response patterns to assign zygosity; twin pairs who indicated high levels of similarity and confusability were more likely to be MZ, and those who indicated the opposite were more likely to be DZ. Thus, MZ twins who reported to have low levels of similarity and confusability (e.g. of ordinary resemblance, never or rarely confused by parents and relatives) would be classified as DZ twins, and DZ twins who reported to have high levels of similarity and confusability (e.g. two peas in a pod, sometimes and always confused by others) would be assigned as MZ pairs. Ultimately, for MZ pairs describing themselves as dissimilar or DZ pairs describing themselves as similar, the misclassification that results is inherent in the data and not a modifiable consequence of the psychometric or classificatory models. Although MZ pairs are, typically, more alike, it is possible that some pairs have distinct differences (e.g. birthmarks or a different haircut) that make them less easily confused.

Among twin pairs with no DNA-based zygosity, 164 pairs (2.6% of the current sample) were assigned different zygosity by the three classification methods. The response patterns of these twin pairs did not reflect those typical of MZ or DZ pairs (Supplementary Figure 6), rendering it difficult to assign consistent zygosity across methods. We plotted the estimated parameters from the three classification methods in Supplementary Figure 7. For twin pairs who were consistently assigned as MZ or DZ pairs, the estimated parameters were highly correlated (all *r*s > .90). However, the associations among the three methods ranged from none to strong for twin pairs who received inconsistent zygosity assignment. Considering that items on the TPQ reflect twin pairs' subjective perception of their similarity and confusability, it is not an infallible method of assigning zygosity in twin pairs who have yet been genotyped.

As studies have suggested that twin pairs, especially DZ pairs, become more different as they age (Beam & Turkheimer, 2013; Scarr & McCartney, 1983), we further explored the extent to which similarity and confusability differ as a function of age among the sample of twins with DNA-based zygosity. To estimate the association between age at the time the questionnaire was completed and the underlying latent trait (θ) of similarity and confusability, age was regressed onto the latent variable of similarity and confusability in the IFA model. Results showed a negative relation between the age and the latent trait of confusability ($r = -.013$ and $-.139$; $p = .758$ and .046 for MZ and DZ pairs, respectively). Although the correlation coefficients were not significantly different (Wald test: $\chi^2(1) = 3.239$, $p = .072$), our findings suggested that similarity

https://doi.org/10.1017/thg.2020.64 Published online by Cambridge University Press

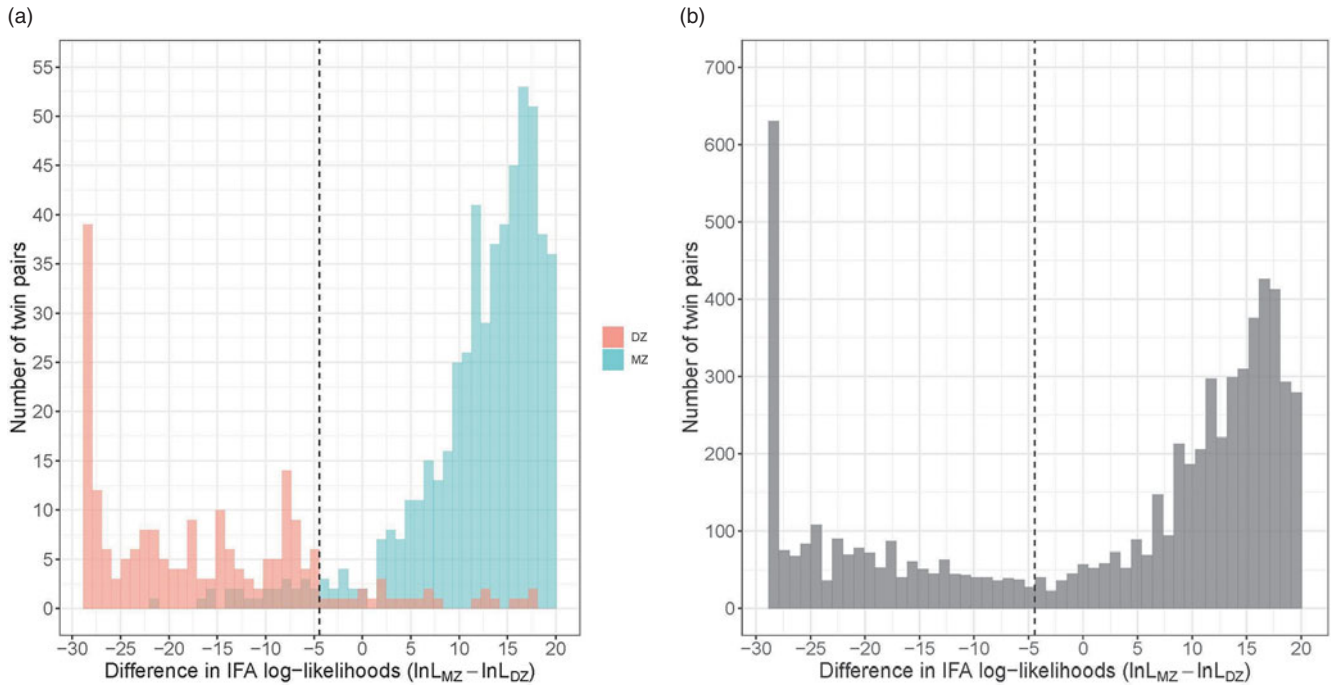

Fig. 3. Distribution of response probabilities from the item factor analysis (IFA) model among twin pairs with and without DNA-based zygosity. Dashed line indicates the optimum cutoff value at $\Delta lnL_{IFA} = -4.45$. (a) Twin pairs with DNA-based zygosity. (b) Twin pairs without DNA-based zygosity.

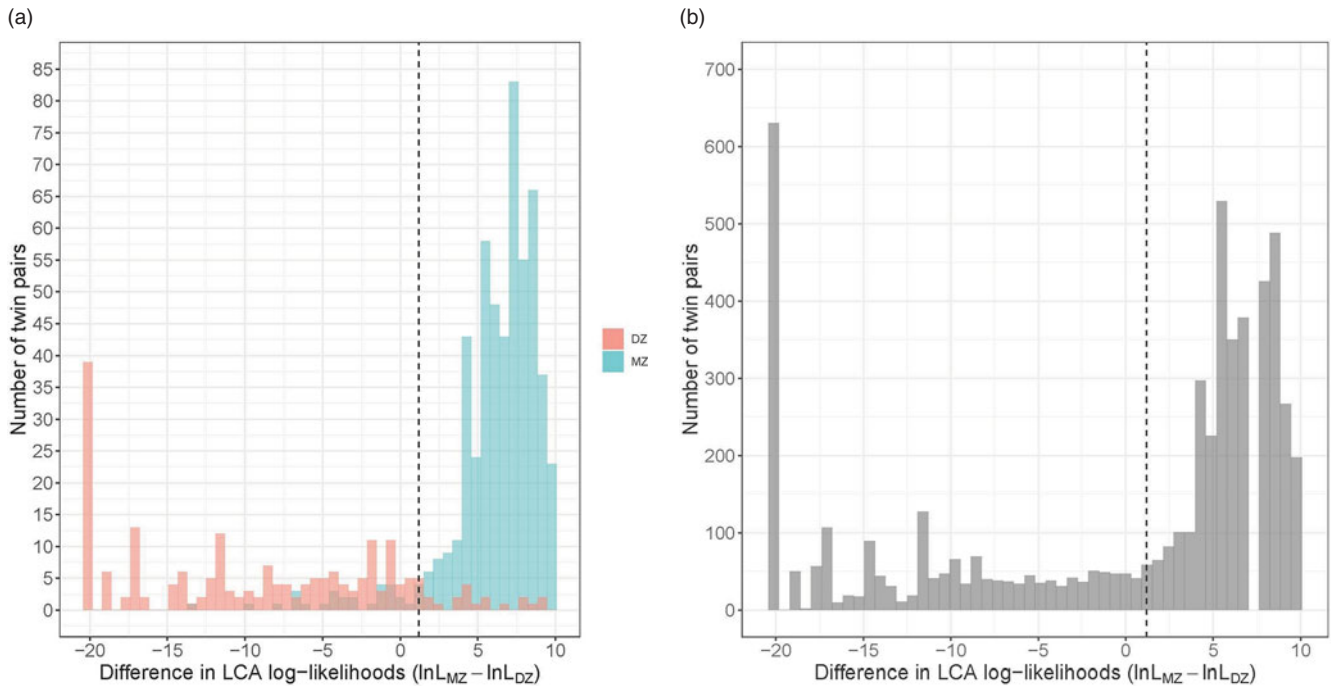

Fig. 4. Distribution of response probabilities from the latent class analysis (LCA) model among twin pairs with and without DNA-based zygosity. Dashed line indicates the optimum cutoff value at $\Delta lnL_{LCA} = 1.2$. (a) Twin pairs with DNA-based zygosity. (b) Twin pairs without DNA-based zygosity.

decreases with age and more so among DZ pairs than MZ pairs. This will be an interesting finding to pursue in future studies, especially in younger twins whose appearance may be changing more rapidly.

Overall Discussion

In this article, we examined the item properties of the TPQ items among a sample of same-sex twin pairs with DNA-based zygosity, and a larger sample of pairs for which DNA-based zygosity was unknown. We evaluated the TPQ both as a psychometric instrument for the measurement of the construct of 'confusability' and as a classification tool for the identification of MZ and DZ pairs. With the exception of the dichotomous 'two peas' item, three of the mistakenness items showed DIF. MZ and DZ twin pairs may differ in their response patterns on these items, even if they endorse similar latent traits of similarity and confusability. Upon examining three

https://doi.org/10.1017/thg.2020.64 Published online by Cambridge University Press

methods to determine zygosity of same-sex twin pairs, we found that the use of unit-weighted PZS scores was sufficient to provide zygosity assignment with high (>90%) overall classification accuracy. The distributions of PZS scores were markedly different in MZ and DZ pairs, not only in their mean but also in their variability and skew. Finally, we conclude that despite the possibilities of misclassification, the TPQ can be regarded as a generally accurate method to determine zygosity among twin pairs who have not been genotyped. The TPQ is somewhat more accurate in the identification of MZ than DZ pairs, for reasons that are inherent in the nature of twin-pair similarity; there are strong limits on the dissimilarity of MZ pairs, whereas DZ pairs can often be highly similar.

A few limitations of this study should be noted. First, the majority of participants in the WSTR self-identified as Caucasian, which may limit the extent to which our findings can be generalized to other racial and ethnic groups. We urge researchers to replicate our findings using data from twin registries with more racial and ethnic diversity. Second, the TPQ was administered upon participants' enrollment to the WSTR. Given the cross-sectional nature of the data, we were unable to examine potential changes in TPQ responses over time (e.g. whether twin pairs are more or less likely to claim similarity as they age), and whether such age-related changes may be larger among DZ than MZ twins. Third, as twin pairs are registered with the WSTR on a volunteer basis, twin pairs who consider themselves to be more similar to one another may be more likely to self-select to participate in twins research. It is possible that DZ twins in the current sample are those who identify as being more alike (i.e. more like twins), which might have biased the likelihood estimates in the current study.

In summary, the TPQ is a generally accurate but by no means infallible method of diagnosing zygosity in twins who have not been genotyped. Even in an era when easier access to DNA has made it possible to diagnose twin zygosity directly without resorting to self-report questionnaires, the ongoing use of large population twin datasets will continue to necessitate the peas questionnaire. Understanding its psychometric and predictive properties will help researchers use this well-worn, yet still useful, tool more effectively.

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/thg.2020.64.

References

Beam, C. R., & Turkheimer, E. (2013). Phenotype-environment correlations in longitudinal twin models. *Development and Psychopathology*, *25*, 7–16.

Buchwald, D., Herrell, R., Ashton, S., Belcourt, M., Schmaling, K., & Goldberg, J. (1999). The Chronic Fatigue Twin Registry: Method of construction, composition, and zygosity assignment. *Twin Research*, *2*, 203–211.

Budowle, B., Moretti, T. R., Baumstark, A. L., Defenbaugh, D. A., & Keys, K. M. (1999). Population data on the thirteen CODIS core short tandem repeat loci in African Americans, U.S. Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians. *Journal of Forensic Sciences*, *44*, 1277–1286.

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. Wiley.

Duncan, G. E., Avery, A. R., Strachan, E., Turkheimer, E., & Tsang, S. (2019). The Washington State Twin Registry: 2019 update. *Twin Research and Human Genetics*, *22*, 788–793.

Eisen, S., Neuman, R., Goldberg, J., Rice, J., & True, W. (1989). Determining zygosity in the Vietnam Era Twin Registry: An approach using questionnaires. *Clinical Genetics*, *35*, 423–432.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*, 378–382.

Forsberg, C. W., Goldberg, J., Sporleder, J., & Smith, N. L. (2010). Determining zygosity in the Vietnam era twin registry: An update. *Twin Research and Human Genetics*, *13*, 461–464.

Hannelius, U., Gherman, L., MäkeläLindstedt, V. A., VLindstedt, A., Zucchelli, M., Lagerberg, C., & Lindgren, C. M. (2007). Large-scale zygosity testing using single nucleotide polymorphisms. *Twin Research and Human Genetics*, *10*, 604–625.

Jackson, R. W., Snieder, H., Davis, H., & Treiber, F. A. (2001). Determination of twin zygosity: A comparison of DNA with various questionnaire indices. *Twin Research*, *4*, 12–18.

Jarrar, Z. A., Ward, K. J., Mangino, M., Cherkas, L. F., Gill, R., Gillham-Nasenya, I., & Spector, T. D. (2018). Definitive zygosity scores in the Peas in the Pod Questionnaire is a sensitive and accurate assessment of the zygosity of adult twins. *Twin Research and Human Genetics*, *21*, 146–154.

McCutcheon, A. C. (1987). *Latent class analysis*. Sage.

Magnus, P., Berg, K., & Nance, W. E. (1983). Predicting zygosity in Norwegian twin pairs born 1915–1960. *Clinical Genetics*, *24*, 103–112.

Magnusson, P. K., Almqvist, C., Rahman, I, Ganna, A, Viktorin, A, Walum, H, & Lichtenstein, P. (2013). The Swedish Twin Registry: Establishment of a biobank and other recent developments. *Twin Research and Human Genetics*, *16*, 317–329.

Merriman, C. (1924). The intellectual resemblance of twins. *Psychological Monographs*, *33*, i57.

Muthén, L. K., & Muthén, B. (2012). *Mplus. Statistical analysis with latent variables. User's guide*. Muthen & Muthen.

Ohm Kyvik, K., & Derom, C. (2006). Data collection on multiple births – establishing twin registers and determining zygosity. *Early Human Development*, *82*, 357–363.

R Development Core Team. (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Reed, T., Plassman, B. L., Tanner, C. M., Dick, D. M., Rinehart, S. A., & Nichols, W. C. (2005). Verification of self-report of zygosity determined via DNA testing in a subset of the NAS-NRC twin registry 40 years later. *Twin Research and Human Genetics*, *8*, 362–367.

Sarna, S., Kaprio, J., Sistonen, P., & Koskenvuo, M. (1978). Diagnosis of twin zygosity by mailed questionnaire. *Human Heredity*, *28*, 241–254.

Scarr, S., & McCartney, K. (1983). How people make their own environments: A theory of genotype greater than environment effects. *Child Development*, *54*, 424–435.

Song, Y. M., Lee, D., Lee, M. K., Lee, K., Lee, H. J., Hong, E. J., Han, B., & Sung, J. (2010). Validity of the zygosity questionnaire and characteristics of zygosity-misdiagnosed twin pairs in the Healthy Twin Study of Korea. *Twin Research and Human Genetics*, *13*, 223–230.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, *91*, 1292–1306.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*, 58–79.

Yang, M. J., Tzeng, C. H., Tseng, J. Y., & Huang, C. Y. (2006). Determination of twin zygosity using a commercially available STR analysis of 15 unlinked loci and the gender-determining marker amelogenin a preliminary report. *Human Reproduction*, *21*, 2175–2179.