Physical Features Observation: Is it Repeatable in Zygosity Determination of Chinese Adult Twins?

Wenjing Gao, Liming Li, Weihua Cao, Siyan Zhan, Yunlong Zhao, Hui Wang and Yonghua Hu

Department of Epidemiology and Biostatistics, School of Public Health, Peking University Health Science Center, Beijing, China

This study reports on the inter- and intrarater reliability of physical features observation. Study subjects were 176 Chinese adult persons, consisting of 89 males and 87 females. Three trained research assistants responded simultaneously and respectively to 12 items regarding the subject's physical features including 'hair', 'Mongoloid folds', left and right 'ear lobes', 'earwax', 'nostril shape', 'tongue rolling', left and right 'hitchhiker's thumb', 'mid-digital hair' and left and right 'simian crease' at the moment of interview. And 14 days later, these subjects received the same observation once again. The results showed that the inter- and intra-observer agreements of 'hair', 'earwax', 'tongue rolling', 'mid-digital hair' and 'simian crease' were almost perfect with most *kappa* (κ) coefficients $\geq .80$, while 'Mongoloid fold' and 'nostril shape' showed poor inter-observer agreement and 'nostril shape' showed poor intra-observer agreement ($\kappa < .40$). Two other physical features, 'hitchhiker's thumb' and 'ear lobes' showed moderate inter-observer agreement and three features, 'hitchhiker's thumb', 'ear lobes' and 'Mongoloid fold', showed moderate intra-observer agreement ($.40 \leq \kappa < .80$). In conclusion, this study suggests that as far as reliability is concerned, the five features which were 'hair', 'earwax', 'tongue rolling', 'mid-digital hair' and 'simian crease' could be considered in zygosity determination of Chinese adult twins, while the two features, 'Mongoloid fold' and 'nostril shape', should be abandoned.

Keywords: twins, zygosity, validation study

Zygosity classification between monozygotic (MZ) and dizygotic (DZ) twins is an essential step for classical twin studies. In the past 20 years, DNA analysis has been considered as the 'gold standard' for zygosity determination, which decreases the level of false classification to be close to zero. However, it is not feasible to apply DNA analysis in large-scale epidemiological studies because it is very time consuming and expensive. In contrast, using a questionnaire to evaluate zygosity based on physical features of twins is simple and useful. Many twin studies have shown that questionnaire-based zygosity diagnosis, in which researchers frequently used such questions as 'Are you as alike as two peas in a pod' or 'Do strangers have difficulty telling you apart', could achieve accuracy of around 95% (Rietveld et al., 2000). Similarly, our previous study using questionnaire and physical features comparison in Chinese adult twins showed 90.1% of MZ and DZ twins could be differentiated correctly (Gao et al., 2006). In this study, we examined 20 physical features and found the agreements between 12 of the physical features and DNA classifications were statistically significant. These features were 'hair', 'Mongoloid folds', left and right 'ear lobes', 'earwax', 'nostril shape', 'tongue rolling', left and right 'hitchhiker's thumb', 'mid-digital hair' and left and right 'simian crease'.

In recent years, a few investigators have been concerned about the reliability of the zygosity questionnaires (Chen et al., 1999; Jackson et al., 2001; Peeters et al., 1998). Jackson showed the test–retest reliabilities of a telephone administered questionnaire interview in 46 mothers. The reliability of the whole questionnaire was .79, and the reliabilities for the two physical features in the questionnaire were only .56 for 'color of hair' and .55 for 'color of eyes' (Jackson et al., 2001). Since most Chinese people have dark hair and dark eyes, we seldom use these two items for zygosity diagnosis. In young Chinese twins (Taiwanese Han population), Chen found that the probability of test-retest and inter-rater agreement of twin similarity questionnaire ranged from 73.7 to 100.0%. The physical features included 'skin color', 'hair texture', 'ear lobe shape', 'hair whorl', 'thumb curvature', 'palmar creases' and 'eyebrow' (Chen et al., 1999). But to date, researchers have concentrated only on the repeatability of self-report and parental-report physical features.

Received 23 March, 2009; accepted 20 November, 2009.

Address for correspondence: Liming Li, Department of Epidemiology and Biostatistics, School of Public Health, Peking University Health Science Center, 38 Xueyuan Road, Haidian District, Beijing 100191, China. E-mail: lmlee@pumc.edu.cn

There has been no report on the reliability of zygosity features by interviewers and no report on the reliability in adult population. Therefore, this study was conducted to evaluate the inter- and intrarater agreement of physical features observation in a Chinese adult population.

Materials and Methods

Study subjects were selected from a general population which is similar in age, gender and ethnicity, to the study population used in a previous twin study evaluating the accuracy of zygosity by questionnaire and physical features comparison (Gao et al., 2006). One hundred and seventy six Chinese adults (89 males and 87 females) were recruited from a community population with their written informed consent. The mean age was 40.36 ± 15.99 years (range = 18 to 80 years). Three trained interviewers responded simultaneously and respectively to 12 items regarding the subject's physical features including 'hair', 'Mongoloid folds', left and right 'ear lobes', 'earwax', 'nostril shape', 'tongue rolling', left and right 'hitchhiker's thumb', 'mid-digital hair' and left and right 'simian crease' at the moment of interview (see Appendix A). To minimize the subjectiveness of answers, the questionnaire asked the three interviewers to describe subjects' physical features according to the standard definitions, for which they had been strictly trained prior to interview. Fourteen days later, 141 adults of these subjects received the same observation once again (66 males and 75 females), with the follow-up rate as high as 80.1%.

Results

Interrater Agreements

Any two of the three interviewers' answers in the same observation were compared, that is, the first interviewer's answers were compared with the second interviewer's answers in the first observation, and the first interviewer's answers were compared with the third interviewer's answers in the first observation, and so on. Cohen's *kappa* coefficients (κ) were calculated to measure the inter-rater agreements (Table 1) (Cohen, 1960).

Table 1 showed that in the two observations, any two of the three investigators had similar interrater reliability for each item. In accordance with the agreement, we could divide these items into three groups: the first group showed the greatest interrater reliability, including 'hair', 'earwax', 'tongue-rolling', 'mid-digital hair' and 'simian crease' with half or more κ coefficients ≥.80; 'hitchhiker's thumb' and 'ear lobe' belonged to the second group, which showed moderate inter-observer agreement with half or more κ coefficients ranged from .40 to .79; 'Mongoloid fold' and 'nostril shape' showed the poorest agreement with half or more κ coefficients below .40.

These interrater reliabilities were similar regardless of subjects' gender and age. In each age and gender category, 'hair', 'earwax', 'tongue-rolling', 'mid-digital hair' and 'simian crease' had good agreements; those with moderate agreements were 'hitchhiker's thumb' and 'ear lobe'; and the poorest agreements were 'Mongoloid fold' and 'nostril shape'. Only the order of some features changed a little ('mid-digital hair' went up in the male subjects, 'nostril shape' went up in the subjects aged below 40, and 'earwax' and 'mid-digital hair' went up in those aged 40 and above) (data not shown).

Intrarater Agreements

Each interviewer's answers in the first observation were compared with those in the second observation, that is, the first and the second and the third interviewer's answers in the first observation were compared with their answers in the second observation respectively. Similarly, Cohen's κ coefficients were calculated to measure the intra-rater agreements (Table 2).

Table 1

Interrater Agreement Among Three Interviewers (*Kappa* Value)[1]

Physical features	1-2-1[2]	1-3-1	2-3-1	1-2-2	1-3-2	2-3-2
Hair	1.000	1.000	1.000	0.977	0.930	0.908
Earwax	0.968	0.984	0.984	1.000	0.957	0.957
Tongue-rolling	0.935	0.987	0.949	0.874	0.954	0.905
Mid-digital hair	0.967	0.899	0.901	0.915	0.913	0.959
Simian crease (L)	0.883	0.883	1.000	0.826	0.850	0.700
Simian crease (R)	0.883	0.833	1.000	1.000	0.789	0.789
Hitchhiker's thumb	0.738	0.704	0.805	0.787	0.745	0.730
Ear lobe (L)	0.441	0.499	0.680	0.554	0.517	0.795
Ear lobe (R)	0.490	0.552	0.663	0.640	0.534	0.687
Mongoloid fold (L)	0.329	0.443	0.318	0.278	0.333	0.263
Mongoloid fold (R)	0.275	0.486	0.308	0.253	0.390	0.432
Nostril shape	0.182	0.158	0.304	0.192	0.276	0.464

Note: [1] The *p* value for each *kappa* is below .05

[2] 1-2-1: the first '1' indicates the first interviewer, the '2' indicates the second interviewer and the second '1' indicates the first observation, and so on.

In the two observations, three investigators showed similar intrarater reliability for each item. The items were divided into three groups: the first group showed the greatest intra-rater reliability, including 'hair', 'tongue-rolling', 'earwax', 'simian crease' and 'mid-digital hair', with two-thirds or more κ coefficients ≥.80; 'hitchhiker's thumb', 'ear lobe' and 'Mongoloid fold' showed moderate intrarater agreement with two-thirds or more κ coefficients ranging from .40 to .79; and 'nostril shape' showed the poorest agreement with two-thirds or more κ coefficients below .40.

Compared with the inter-rater reliability, the intra-rater reliability showed almost the same situation except 'tongue-rolling' and 'simian crease' ranking went up and 'Mongoloid fold' went into the second group with moderate agreement.

Almost the same results were found in the male and female subjects and those aged 40 and above and aged below 40, respectively (data not shown). The physical features with good intra-rater agreements were still 'hair', 'earwax', 'tongue-rolling', 'simian crease' and 'mid-digital hair', and the poorest agreements were 'Mongoloid fold' and 'nostril shape'. Just like the inter-rater agreement, the order of some features changed a little in some subgroups.

Discussion

Cohen's κ coefficient is a statistical measure of inter-rater and intra-rater agreement (Cohen, 1960). It is generally thought to be a more robust measure than simple percentage agreement calculation since *kappa* takes into account the agreement occurring by chance. According to the interpretation of *kappa* values by Landis and Koch (1977), < 0 means no agreement, .00–.19 means poor agreement, .20–.39 means fair agreement, .40–.59 means moderate agreement, .60–.79 means substantial agreement, and .80–1.00 means almost perfect agreement. However, in this study the *kappa* values were categorized into three groups, good agreement (≥.80), moderate (.40–.79) and poor (<.40), which is more strict than the traditional interpretation and fits the data here better.

It has been noted that the number of categories and subjects will affect the magnitude of the value. The *kappa* value will be higher when there are fewer categories. In this study, each item had the same number of categories and all subjects responded to all items, which consequently did not affect the difference of *kappa* values among different features.

Interrater Reliability

Five items ('hair', 'earwax', 'tongue-rolling', 'mid-digital hair', 'simian crease') were found to have almost perfect inter-observer reliability (κ ≥.80) in the two surveys. Observations on 'hair' and 'earwax' could not simply rely on on-site observation since people could make their hair straight or curly as they please and not every person had earwax available. In this study, three trained investigators responded simultaneously and respectively to the items; that is, as long as one investigator asked the subjects whether he or she had straight/curly hair or dry/wet earwax, the other two investigators wrote down the answers at the same time. As a result, it was expected that these two indicators should have a perfect agreement among three investigators and in theory, all the *kappa* values should be 1.00, although this was not the case. It may be partly due to some errors in the investigators' recording. For the three items 'mid-digital hair', 'simian crease' and 'tongue-rolling', there were significant differences between dominant and recessive individuals, which resulted in better investigator consistency. In addition, only a few subjects had a simian crease (8.0% in the first investigation of the first investigator), which to some extent led to the high *kappa* value of this item. This meant that, regardless of good reliability, 'simian crease' could play a limited role in distinguishing between MZ and DZ twins. For

Table 2

Intrarater Agreement of Three Interviewers (Kappa Value)[1]

Physical features	Interviewer 1	Interviewer2	Interviewer3
Hair	1.000	1.000	0.884
Tongue-rolling	0.987	0.949	0.830
Earwax	0.984	0.984	0.770
Simian crease (L)	0.883	1.000	0.758
Simian crease (R)	0.833	1.000	0.930
Mid-digital hair	0.899	0.901	0.590
Hitchhiker's thumb	0.704	0.805	0.602
Ear lobe (L)	0.499	0.680	0.711
Ear lobe (R)	0.552	0.663	0.626
Mongoloid fold (L)	0.443	0.318	0.433
Mongoloid fold (R)	0.486	0.308	0.503
Nostril shape	0.158	0.304	0.424

Note: [1]The *p* value for each *kappa* is below .05.

the item 'mid-digital hair', the investigator should pay special attention to environmental factors, because some subjects' fingers had been hurt; Indeed some said they had mid-digital hair in their childhood, but after working for a long time, it had been worn away. For 'tongue-rolling', some subjects could roll the tongue into a tube-like shape, but they needed to practice before the official evaluation. The duration of the practice would directly affect the inter-rater agreements. In this study, two minutes was allocated for tongue-rolling practice. In order to avoid influence of others, the investigators allowed the subjects to show the tongue-rolling without onlookers.

Two items ('hitchhiker's thumb' and 'ear lobe') were found to have moderate or substantial inter-observer reliability ($.80 > \kappa \geq .40$) in the two surveys. There are similarities in the definition of the two features — they are both quantitative traits. For hitchhiker's thumb, homozygous recessives can bend the distal joint of the thumb backward to a nearly 90° angle; the heterozygous or homozygous dominant condition yields thumbs that cannot bend backward more than approximately 30°. Ear lobes may be either adherent or free and pendulous. Homozygous recessives have attached ear lobes with a right or obtuse angle (≥90°) between the cheek and the lower edge of the ear; heterozygous or homozygous dominant individuals have detached (free) ear lobes (<90°). When subjects could bend their thumb nearly 30° or the angle between their cheek and lower edge of ear was nearly 90°, there would be more or less difference among the three investigators' recording. So the interrater reliability of the two items was less than the first five items. In the current study, the item of 'ear lobes' was found to be affected by the environment. If subjects were wearing earrings, then it was more possible that he or she had free ear lobes because of the gravitational force of the earrings (100% of subjects wearing earrings had free ear lobes in this study).

The worst interrater reliability was found to be in the two items 'Mongoloid fold' and 'nostril shape'. This was because the definition of the 'Mongoloid fold' was 'a skin fold of the upper eyelid (from the nose to the inner side of the eyebrow) covering the inner corner of the eye', in which there was no distinct cut-off point between 'yes' and 'no'. During the training, we provided pictures of typical examples of persons with and without Mongoloid fold, but in the survey, not all subjects' traits were as typical as those of the examples. For the nostril shape, the quantitative character made the agreements less satisfactory, which was a similar situation to the observation of 'hitchhiker's thumb' and 'ear lobe'. We defined the broad nostril shape as the angle < 45° between the maximum diameter of the nostril and the horizontal line, the narrow shape with an angle ≥45°. Besides, one's facial expression, such as smiling, breathing heavily, and dilating the nostril intentionally, would all change the shape of the nostrils. In this study, it was occasionally found that when two investigators observed the same subject at different times, the results might have been different just because of the different facial expression of the subjects.

Intra-Observer Reliability

The intrarater reliability showed almost the same story as interrater reliability. The almost perfect agreements were found in the five items 'hair', 'earwax', 'tongue-rolling', 'mid-digital hair' and 'simian crease'. For the two items which were investigated by direct questions, intra-observer agreement for 'hair' was better than that for 'earwax', which showed that people paid more attention to their hair. About a quarter of subjects said they did not notice whether their earwax was dry or wet. However, for 'tongue-rolling', the process of practice and learning in the first observation allowed subjects to roll the tongue more easily during the second test. As a result, this item showed greater intra-observer agreements than that among different observers.

The two quantitative traits 'ear lobe' and 'hitchhiker's thumb' showed poor intra-rater reliability, which was the same as their inter-rater reliability. However, the intra-rater reliability was better for 'Mongoloid fold', which meant for this item, the results of the same investigator were a little more stable regardless of the poor agreement among different investigators. In this context, for the twins' physical features, one investigator was required to observe both twins at the same time.

The item with the poorest agreement was 'nostril shape', the possible reasons for which have been mentioned previously in the section of the inter-rater reliability.

Some researchers found that there were slight age- and sex-differences in the validity of questionnaire-based zygosity in twins (Christiansen et al., 2003). However, in this study, after stratifying by age and sex, all the inter-rater agreements and intra-rater agreements in each stratum showed almost the same situation as that of the total subjects, except some *kappa* values fluctuated, also, some p values were greater than .05 due to the reduced sample size. This showed that age and sex did not have much influence on the reliability of the observation of all physical features.

In conclusion, based on the findings in both inter-rater reliability and intra-rater reliability in a Chinese adult population, the five features 'hair', 'earwax', 'tongue rolling', 'mid-digital hair' and 'simian crease' showed the best reliability and could be considered in zygosity determination of Chinese adult twins, while the two features, 'Mongoloid fold' and 'nostril shape', should be abandoned or reformed before use.

Acknowledgments

This study was supported by grants from China Medical Board (01-746) and Youth Fund of the School of Public Health, Peking University (2006–2007). We thank Mr Zhiping Yu for help in data collection and Dr Chunhui Wang and Mr Ken Rouse during article revision.

References

Chen, W. J., Chang, H. W., Wu, M. Z., Lin, C. C. H., Chang, C., Chiu, Y. N., & Soong, W. T. (1999). Diagnosis of zygosity by questionnaire and polymarker polymerase chain reaction in young twins. *Behavior Genetics, 29*, 115–123.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.

Gao, W., Li, L., Cao, W., Zhan, S., Lv, J., Qin, Y., Pang, Z., Wang, S., Chen, W., Chen, R., & Hu, Y. (2006). Determination of zygosity by questionnaire and physical features comparison in Chinese adult twins. *Twin Research and Human Genetics, 9*, 266–271.

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.

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

Christiansen, L., Frederiksen, H., Schousboe, K., Skytthe, A., von Wurmb-Schwark, N., Christensen, K., & Kyvik, K. (2003). Age- and sex-differences in the validity of questionnaire-based zygosity in twins. *Twin Research, 6*, 275–278.

Peeters, H., Gestel, S. V., Vlietinck, R., Derom, C., & Derom, R. (1998). Validation of a telephone zygosity questionnaire in twins of known zygosity. *Behavior Genetics, 28*, 159–163.

Rietveld, M. J. H., Valk, J. C. V. D., Bongers, I. L., Stroet, T. M., Slagboom, P. E., & Boomsma, D.I. (2000). Zygosity diagnosis in young twins by parental report. *Twin Research, 3*, 134–141.

Appendix A

List of Physical Features

1. Hair	__straight hair	__curly hair	__ hard to say
2. Mongoloid fold[1]			
left	__ present	__ absent	__ hard to say
right	__ present	__ absent	__ hard to say
3. Ear lobes[2]			
left	__ attached	__ detached	__ hard to say
right	__ attached	__ detached	__ hard to say
4. Earwax	__ dry	__ sticky	__ hard to say
5. Nostril shape	__ broad	__ narrow	__ hard to say
6. Tongue rolling[3]	__ yes	__ no	__ hard to say
7. Hitchhiker's thumb[4]	__ present	__ absent	__ hard to say
8. Mid-digital hair[5]	__ present	__ absent	__ hard to say
9. Simian crease[6]			
left	__ present	__ absent	__ hard to say
right	__ present	__ absent	__ hard to say

Note: [1] **Mongoloid fold** (epicanthic fold): This is a skin fold of the upper eyelid (from the nose to the inner side of the eyebrow) covering the inner corner (medial canthus) of the eye. Dominant allele causes it.

[2] **Ear lobes** may be either adherent or free and pendulous. Homozygous recessives have attached ear lobes; heterozygous or homozygous dominant individuals have detached (free) ear lobes.

[3] **Tongue rolling:** Persons with a dominant allele in heterozygous or homozygous condition can roll their tongues into a tube-like shape; homozygous recessives are nonrollers and can never learn to roll their tongues.

[4] **Hitchhiker's thumb**: Homozygous recessives can bend the distal joint of the thumb backward to a nearly 90° angle; heterozygous or homozygous dominant condition yields thumbs that cannot bend backward more than approximately 30°.

[5] **Mid-digital hair:** People lacking hair in the middle segments of the fingers are homozygous recessive. The presence of hair on one or more middle segments of the fingers may be governed by a series of alleles each of which is dominant to the recessive.

[6] **Simian crease**: A simian crease is a single palmar crease as compared to two creases in a normal palm.