Acta Psychologica 230 (2022) 103734

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/locate/actpsy

The House-Tree-Person test is not valid for the prediction of mental health: An empirical study using deep neural networks

Yijing Lin (林依静)[a,1], Nan Zhang (张楠)[a,1], Yukun Qu (瞿宇堃)[b], Tian Li (李添)[a], Jia Liu (刘嘉)[c], Yiying Song (宋宜颖)[a,*]

[a] Beijing Key Laboratory of Applied Experimental Psychology, Faculty of Psychology, Beijing Normal University, Beijing, China
[b] State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing, China
[c] Department of Psychology & Tsinghua Laboratory of Brain and Intelligence, Tsinghua University, Beijing, China

ARTICLE INFO

Keywords:
House-Tree-Person test
Projective drawing test
Mental health
Depression
Deep neural network

ABSTRACT

As one of the projective drawing techniques, the House-Tree-Person test (HTP) has been widely used in psychological counseling. However, its validity in diagnosing mental health problems remains controversial. Here, we adopted two approaches to examine the validity of HTP in diagnosing mental health problems objectively. First, we summarized the diagnostic features reported in previous HTP studies and found no reliable association between the existing HTP indicators and mental health problems studied. Next, after obtaining HTP drawings and depression scores from 4196 Chinese children and adolescents (1890 females), we used the Deep Neural Networks (DNNs) to explore implicit features from entire HTP drawings that might have been missed in previous studies. We found that although the DNNs successfully learned to extract critical features of houses, trees, and persons in HTP drawings for object classification, it failed to classify the drawings of depressive individuals from those of non-depressive individuals. Taken together, our study casts doubts on the validity of the HTP in diagnosing mental health problems, and provides a practical paradigm of examining the validity of projective tests with deep learning.

1. Introduction

The projective drawing test, a technique that interpreted a person's feelings through drawing, has been widely used in psychological counseling for decades (Blatt, 1975; Catterall & Ibbotson, 2000; Garb et al., 2000; Goodenough, 1928; Goodenough & Harris, 1950; Lilienfeld et al., 2000). Because of its lower requirements for language expression ability, it is especially useful for children and individuals with language impairments. In psychological counseling, many kinds of projective drawing tests have been proposed. The House-Tree-Person (HTP) test, one of the most widely-used projective drawing tests, was first put forward by Buck in 1948 (Buck, 1948), based on the principle that drawing can reflect people's inner thoughts. Then it developed into the Synthetic HTP test commonly used in counseling today. In the testing process, the participant is asked to draw a painting containing the house, tree, and person elements on a piece of white paper, and the counselor uses the drawing to analyze the participant's mental state, emotion, personality, etc. Afterwards, the HTP has evolved to be more standardized, with an indicator system composed of critical features in HTP drawings that can predict mental health, such as whether there is a chimney on the roof, whether there is a door on the house, and whether there are scars on the trunk. Some studies suggest that there are qualitative relationships between the HTP indicators and many mental health problems (Garb et al., 2002; Xiang et al., 2020; Xie & Ye, 1994; Yan et al., 2013; Yan et al., 2014). Among these mental health problems, the relationship between HTP indicators and depression has been frequently researched. For example, studies have shown that depression patients draw more blacking and shadows in their drawings than typical individuals (Buck, 1948; Ning et al., 2015).

Although the indicator system has largely regulated the analysis of HTP drawings, there are some limitations in applying the indicator system. First, most indicators do not have operational definitions, leading to the ambiguity of the coding criterion. For example, it is difficult to define whether a picture is "artistic" (Xie & Ye, 1994).

[*] Corresponding author at: Beijing Key Laboratory of Applied Experimental Psychology, Faculty of Psychology, Beijing Normal University, Beijing 100875, China.
E-mail addresses: yijinglin@mail.bnu.edu.cn (Y. Lin), nan_zhang@mail.bnu.edu.cn (N. Zhang), yukun.qu@mail.bnu.edu.cn (Y. Qu), litian@bnu.edu.cn (T. Li), liujiaTHU@tsinghua.edu.cn (J. Liu), songyiying@bnu.edu.cn (Y. Song).

[1] Yijing Lin and Nan Zhang contributed equally to this work and should be considered co-first authors.

https://doi.org/10.1016/j.actpsy.2022.103734
Received 3 March 2022; Received in revised form 8 August 2022; Accepted 31 August 2022
Available online 1 September 2022

Second, different researchers proposed different indicators and interpreted the same indicators differently based on their own experiences.

Moreover, many researchers have doubted the validity of the HTP test (for a review, see Lilienfeld et al., 2000). For example, Joiner et al. (1996) found that none of the studied HTP indicators was significantly related to self-report depression or anxiety. Another study found that adding the HTP test to other psychometric test decreased the accuracy of classifying psychiatric patients from nurses (Wildman and Wildman, 1975). These negative results may be due to the subjectivity of expert judgements and coding process in projective drawing tests. However, no study has systematically and objectively explored whether the HTP test really contain information that can be used to predict mental health. In this study, we adopted two empirical approaches to explore whether there was a reliable relationship between HTP drawings and mental health and provided new empirical evidence questioning the validity of the HTP test.

First, to examine whether existing HTP indicators have a stable relationship with mental health, we systematically summarized and integrated all the indicators reported to be related to mental health problems in the literature and made a comprehensive summary of previous results with a meta-analysis.

Second, we extended from the existing indicators to the features of the entire HTP drawings that might have been missed in previous studies. The meta-analysis was based on the existing indicators proposed by previous researchers and experts. However, there is still the possibility that the features related to mental health are not limited to the existing indicators; some features neglected by previous studies, even some implicit and high-dimensional features in the HTP drawings, may also be useful for diagnosing mental health problems. To rule out this possibility, we used the deep neural network (DNN) to extract potential features from HTP drawings. DNN is a new image classification approach widely used in a variety of fields in recent years (Sladojevic et al., 2016; Sun et al., 2015; Szegedy et al., 2013). With the multi-layer network structure and the training of large number of images, the DNNs can learn to extract features from images and use these features to classify images into different categories (Ciregan et al., 2012; Krizhevsky et al., 2012; Russakovsky et al., 2015). In our study, the DNNs enable us to classify depression and typical participants with their HTP drawings in an objective data-driven way. Here we focused on depression because it is the most frequently studied mental health problem in the meta-analysis.

2. Methods

2.1. Study search and selection

The process of study search and selection was shown in Fig. 1. A systematic article search was carried out in Web of Science and CNKI (China National Knowledge Infrastructure), covering the period from 1947, the year that HTP was first proposed, to 2020. The search was based on three keywords: House Tree Person test (or HTP test), Kinetic-House-Tree-Person test (or KHTP test), and Synthetic House-Tree-Person test, and we browsed all articles and manually selected the articles on mental health. The search was restricted to peer-reviewed studies and master's thesis in English and Chinese. Additionally, the references from the articles we retrieved were cross-checked to verify that all the studies were included and not duplicated.

Two reviewers selected potentially eligible studies. First, we checked the titles and abstracts to ensure that the studies were on the relationship between HTP and mental health. Then, to confirm whether quantitative results on the relationships were reported in the studies, we checked the main body of the articles. Selections were based on a set of inclusion criteria. All disagreements were resolved by consensus.

The inclusion criteria were as follows:

- "Mental health" quotes the definition given by Centers for Disease Control and Prevention in USA, that is, mental health includes our emotional, psychological, and social well-being (World Health Organization, 2004, 2017). Therefore, not only mental disorders but also behavioral tendencies that may be related to mental health were included.

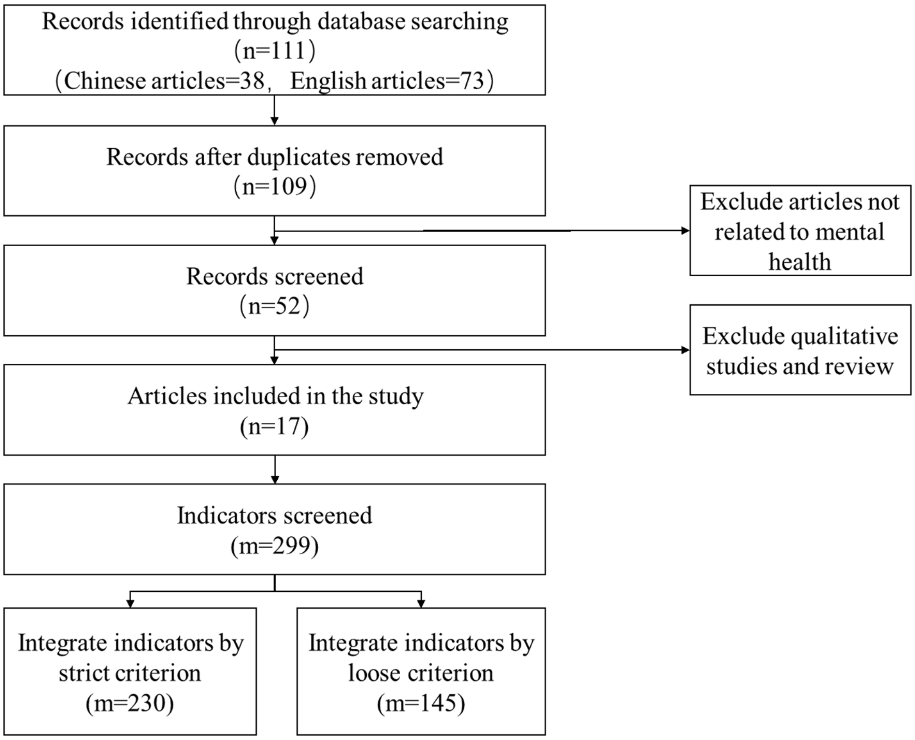
Records identified through database searching
(n=111)
(Chinese articles=38， English articles=73)
Records after duplicates removed
(n=109)
Exclude articles not related to mental health
Records screened
(n=52)
Exclude qualitative studies and review
Articles included in the study
(n=17)
Indicators screened
(m=299)
Integrate indicators by strict criterion
(m=230)
Integrate indicators by loose criterion
(m=145)

Fig. 1. A flow diagram of article searching, article exclusion, and HTP indicator integration.

- Mental health problems were measured by experts or authoritative scales.
- The studies should be quantitative, that is, quantitative results on relationships between HTP indicators and mental health with statistical tests were reported.

2.2. *Indicator extraction and integration*

After study selection, all the indicators in the articles were extracted. The number of indicators proposed and eventually reached statistically significance in every article was recorded. Some studies reported the results of more than one statistical test (e.g., regression analysis, t-test, chi-square test, etc.). Since most studies reported the results of significant test of difference (i.e., chi-square test, t-test), if both regression analysis and significance test of difference were used in the article, we chose to record the latter one.

To integrate indicators across different studies, all the indicators were merged by one reviewer and independently checked by a second reviewer. The indicators were integrated with two criteria. One criterion was strict, that is, only indicators with similar meanings can be merged (e.g., “large house” and “large proportion of house” in the drawing). The other criterion was loose, and indicators with different directions or degrees regarding the same dimension were merged (e.g., “large house” and “small house” were merged into “house size”; “no expression”, “sad expression”, and “dull expression” were merged into “expression”). All disagreements were resolved by consensus.

2.3. *Participants*

We recruited 4598 Chinese children and adolescents ranging from 9 to 18 years old (from the 4th grade of primary school to the 3rd grade of high school). After excluding participants with missing values or poor quality HTP drawings, the final sample size was 4196 (505 from primary school, 2166 from middle school, 1525 from high school; 1890 female). The experiment protocol was approved by the Institutional Review Board. Written informed consent was obtained from all participants before data collection.

2.4. *Assessment of depression*

The Children’s Depression Inventory (CDI) (Kovacs, 1985) was used to measure the depression level of the participants, which was shown to be reliable and valid in Chinese participants (Wu et al., 2010). CDI contained 27 items, and the score of each item ranged from 1 to 3 points. The total score ranged from 27 to 81 points, and a higher score indicated a higher depression level. We excluded participants with 5 and more missing values and replaced other missing values with the average score of all items (Little & Rubin, 2014).

For the classification task of DNN, the depression and typical samples were labeled with two criteria (Fig. 2). In previous studies (Wu et al., 2010), 46 points was used as the boundary to identify depressive individuals. So, we first labeled the sample with the bisection criterion, in which scores of 46 points and above were labeled as depression, and the rest was labeled as typical ($N_{\text{depression}} = 1380$, $N_{\text{typical}} = 2809$). Then, in order to balance the sample size of the depression and typical samples and increase the distinctiveness between them, we also used a trisection criterion, in which the participants were divided into three equal groups. We labeled the group with highest scores (depression score ≥ 47) as depression, and the group with the lowest scores (depression score ≤ 37) as typical ($N_{\text{depression}} = 1230$, $N_{\text{typical}} = 1314$). The third group with medium scores ($37 <$ depression score < 47, $N = 1645$) was not used in depression classification.

2.5. *House-Tree-Person dataset*

The participants were instructed to draw a picture containing three elements, house, tree, and person, and besides these, drawing anything was allowed. They were told that it was not an art examination so that they didn’t need to concern about being scored. An exemplar of our participants’ drawing was shown in Fig. 3a.

After data collection, we scanned the drawings and imported them to the computer. The size of each image was about 3300 × 2400. To improve the quality of the pictures and maximize useful information as much as possible, we cropped the blank borders programmatically and increased image contrast. Then the drawings were resized to 224 × 224 pixels to fit the input size of DNNs.

2.6. *Training AlexNet and InceptionV3 to classify depression*

We first used the AlexNet (Krizhevsky et al., 2012) to extract features from the HTP drawings, which was pre-trained to classify objects in millions of colorful natural images of the ImageNet (Fig. 3b, left) (Russakovsky et al., 2015) into 1000 categories. The AlexNet includes 8 layers with a hierarchical architecture; the first 5 layers are convolutional layers and the last 3 layers are fully-connected layers. In our study, we trained the AlexNet for depression classification using transfer

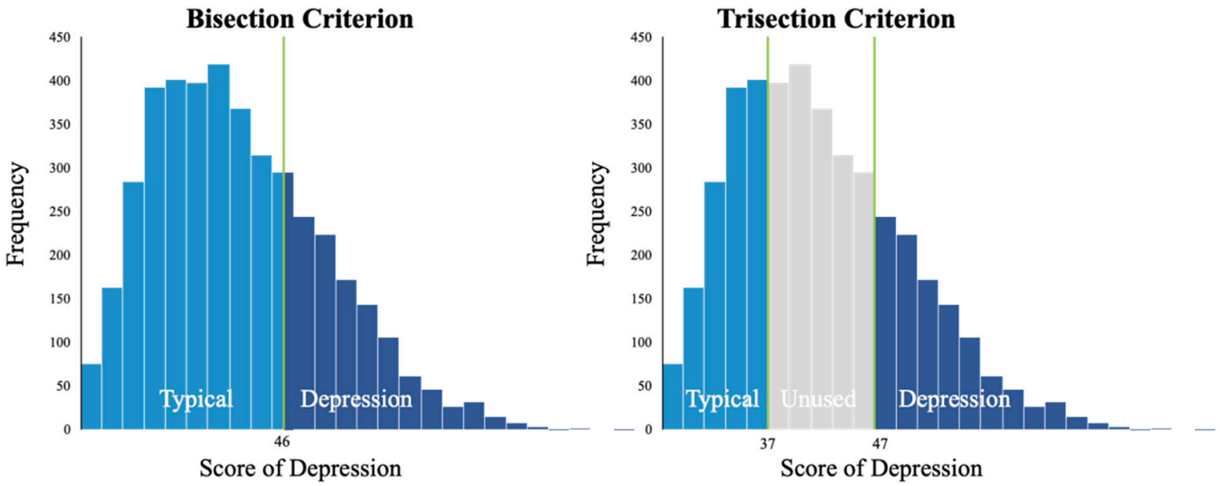

Fig. 2. Two criteria for labeling the depression and typical samples.
Note. The histogram showed the distribution of the scores measured by Children’s Depression Inventory. The left figure represented the bisection criterion, with scores of 46 points and above labeled as depression and the rest labeled as typical ($N_{\text{depression}} = 1380$, $N_{\text{typical}} = 2809$). The right figure illustrated the trisection criterion, with scores of 47 points and above labeled as depression, and scores of 37 points and below labeled as typical ($N_{\text{depression}} = 1230$, $N_{\text{typical}} = 1314$).

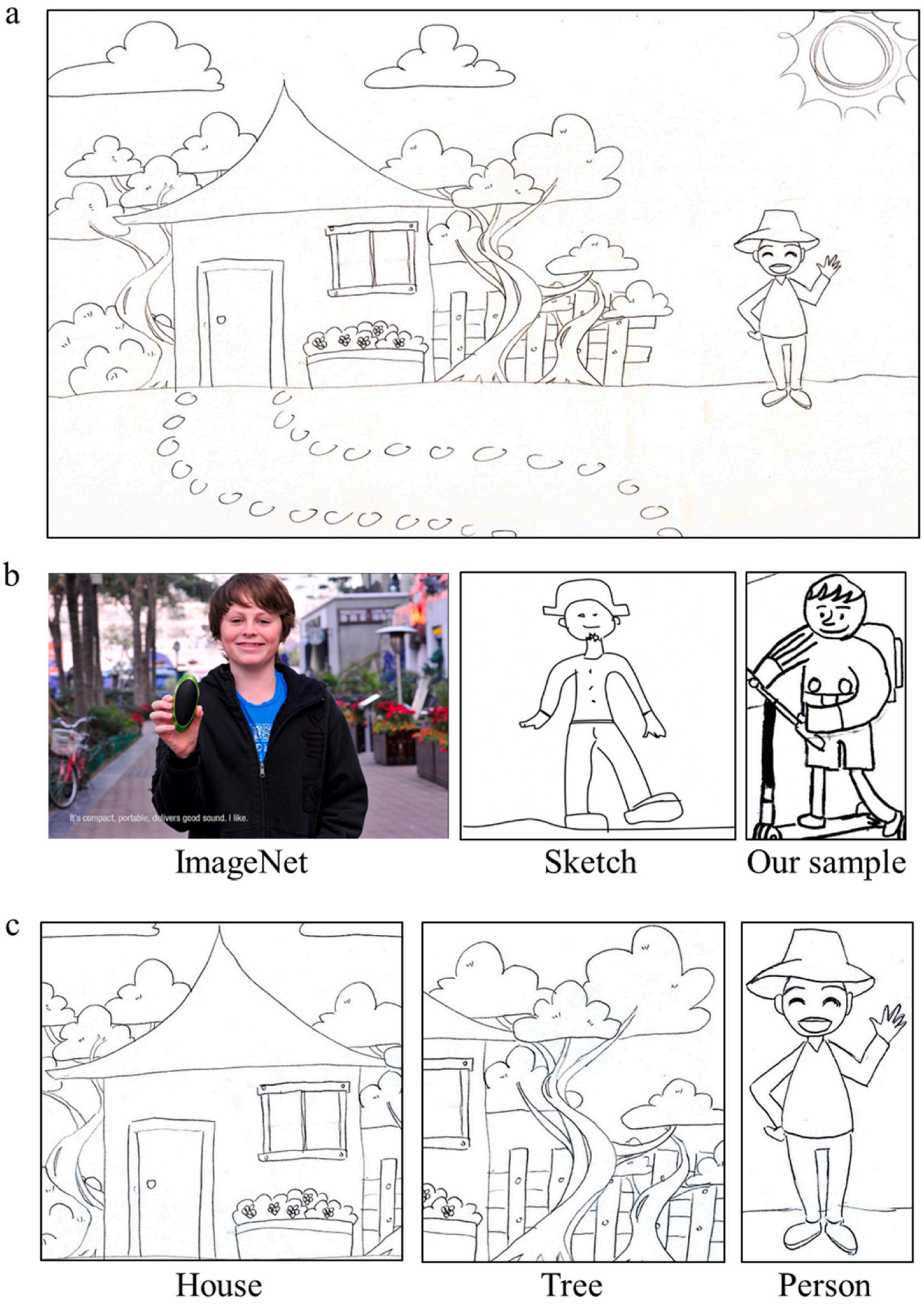

Fig. 3. The House-Tree-Person drawing and the images of different pre-training datasets
Note. a) An exemplar HTP drawing of our participants. b) The images from different pre-training datasets. The images from ImageNet had color and texture, while the images from Sketch had only contour information, which were more similar to our samples. c) To generate the images of houses, trees and persons, we manually marked and cut the typical houses, trees, and persons from the whole HTP drawings.

learning. Specifically, we replaced the final layer of the AlexNet with a two-unit layer for the classification of depression and non-depression samples and froze all the parameters except the connection to the final layer in transfer learning. We randomly selected 90 % of the images from the depression and typical samples to constitute the training dataset, and the rest 10 % images formed the test dataset. With the bisection criterion, the training dataset consisted of 1242 depression images and 2527 typical images, and the test dataset consisted of 138 depression images and 282 typical images. With the trisection criterion, the training dataset consisted of 1107 depression images and 1182 typical images, and the test dataset consisted of 123 depression images and 132 typical images. We used the Python package DNNbrain (Chen et al., 2020) to train the network. The loss function was cross-entropy, and the optimizer was Adam. The learning rate was 0.01, and the network was trained for 60 epochs.

The test results were described with accuracy, precision, recall, and F1 score. The accuracy was defined as the ratio of the correct labeled samples to all the given samples. The precision was defined as the proportion of the samples correctly judged as depression to all the samples judged as depression, and the recall was defined as the proportion of the samples correctly judged as depression to all the real depression samples. Because there is a trade-off between precision and recall, we also used the F1 score, calculated as 2 times precision times recall divided by the sum of precision and recall. The closer the F1 score is to 1, the better the DNN classifies depression.

In addition, we also used the InceptionV3 for classifying depression, which has more complex network architecture and higher pre-training performance than the AlexNet (Szegedy et al., 2016). InceptionV3 includes 5 convolutional layers, 10 Inception modules, and 1 fully-connected layer. Each Inception module consists of several convolutional layers with different kernel sizes, which makes fully use of the image features in different scales. The same procedures of transfer learning were applied for InceptionV3.

2.7. Pre-training Sketch-A-Net to classify sketch images

In order to make the pre-training images more similar to the HTP drawings used in transfer learning, we changed the pre-training dataset from colorful natural images (Fig. 3b, left) to black and white sketch images. We first pre-trained the DNN to classify sketch objects. To this end, we manually marked and cut the typical houses, trees, and persons from the HTP drawings to generate the House, Tree, and Person samples (Fig. 3c). If there was more than one house/tree/person in the drawing, the houses/trees/humans with the same configurations were marked as one sample, while those with different forms and configurations were marked as different samples; therefore, one HTP drawing may generate more than one sample for each category. Otherwise, as in Fig. 3a, the trees were similar to each other, so we only chose one as a representative. In the pre-training dataset, we only used the drawings from the participants with middle level of depression scores ($37 <$ depression score < 47), so that the pre-training samples for object classification and training samples for depression classification would be similar but not overlapping.

Next, to enlarge the pre-training dataset, we combined Sketchy (Sangkloy et al., 2016) and Tu-Berlin (Eitz et al., 2012) to form the Sketch dataset (Fig. 3b, middle), which contain over 100,000 sketch images in 291 categories. Then, we replaced the person category in the Sketch dataset with our Person samples (Fig. 3b, right) and added the house and tree categories with our House and Tree samples. Altogether, there were 293 categories. Finally, the pre-training dataset was randomly divided into a training dataset (90 % images of each category) and a validation dataset (10 % images of each category). The total size of the training dataset was 95,475 (House = 1641, Tree = 1679, Person = 1731, other categories = 90,424, with each category ranging from 70 to 700). And the total size of the validation dataset is 10,620 (House = 183, Tree = 187, Person = 192, other categories = 10,058).

We pre-trained the DNN to classify sketch images into 293 object categories. Here we used Sketch-A-Net, which has been used in classification of sketch images (Yu et al., 2017). Compared with AlexNet, Sketch-A-Net only differed in the bigger kernel size of the first convolution layer, so that it can make more use of the information of the whole picture. The training parameters were the same as those for AlexNet and InceptionV3, except that the number of epochs was 30.

After the pre-training, to test whether this pre-trained Sketch-A-Net can successfully classify objects used in the following depression classification, the pre-trained network was tested to classify houses/trees/persons from a test dataset drawn by depressive (depression score ≥ 47) and typical participants (depression score ≤ 37), including 2783 houses, 2867 trees, and 3017 persons.

2.8. Training the pre-trained Sketch-A-Net to classify depression

We then trained the pre-trained Sketch-A-Net for depression classification using the same transfer learning procedures and HTP datasets labeled with the trisection criteria as for AlexNet and InceptionV3. Additionally, depression classification using the Houses, Trees, and Persons samples cut from HTP drawings was also tested. The House dataset included 2503 ($N_{\text{depression}} = 1211$, $N_{\text{typical}} = 1292$) training images and 278 ($N_{\text{depression}} = 134$, $N_{\text{typical}} = 144$) test images; the Tree dataset included 2578 ($N_{\text{depression}} = 1260$, $N_{\text{typical}} = 1318$) training images and 287 ($N_{\text{depression}} = 140$, $N_{\text{typical}} = 147$) test images; the Person dataset included 2712 ($N_{\text{depression}} = 1301$, $N_{\text{typical}} = 1411$) training images and 302 ($N_{\text{depression}} = 145$, $N_{\text{typical}} = 157$) test images. Materials and analysis code for this study are available by emailing the corresponding author. This study was not preregistered.

3. Results

3.1. Selection of previous HTP studies

To examine whether HTP indicators have a stable relationship with mental health, we first searched for all the previous studies on the relationship and summarized all the diagnostic HTP indicators reported to be related to mental health. As shown in Fig. 1, 38 Chinese articles and 73 English articles were retrieved from the databases. After reviewing the titles and abstracts and removing duplicates, 53 of 109 articles related to mental health were identified as potentially eligible. After checking whether quantitative results on relationships between HTP indicators and mental health with statistical tests were reported in the full text, a total of 17 articles were finally included.

The characteristics of each included article were described in Table 1. As shown in the table, 7 of the 17 articles studied the relationship between HTP indicators and depression; 2 articles studied anxiety; 2 articles studied schizophrenia; 2 articles studied somatization; 1 article studied anxiety and post-traumatic stress disorder. The remaining 3 articles studied dependent personality disorder, suicide, and high-functioning autism, respectively. As for the statistical tests, 10 of the 17 articles reported the results of chi-square tests, 3 articles reported the results of regression analysis, 2 articles reported the results of correlation analysis, and 1 article reported the results of t-test.

3.2. Indicator extraction and integration

The number of indicators proposed in each study and those reaching statistical significance were summarized in Table 1 and Fig. 4. Across all studies, an average of 18 significant indicators was reported in each article, and 11 of the 17 articles reported <15 significant indicators. The proportions of significant indicators in different articles varied widely, from 3.6 % to 96 %. Among the 5 articles proposing >100 indicators, the proportion of significant indicators did not reach 30 %, and >70 % of the proposed indicators proved to be invalid. For the remaining 11 articles with less proposed indicators, 7 papers reported <10 significant indicators. In the article with the highest proportion of significant indicators (96 %) (Xie & Ye, 1994), only 25 indicators were proposed. Finally, there was one article that didn't report the number of proposed indicators, and the number of significant indicators was 21 (Qiu & Wu, 2010). In short, these results indicated that although there were as many as several hundred indicators in the HTP manual, only a small proportion of them was reported to be significantly related to mental health.

Because different studies used different indicator systems and there was an overlapping and nested relationship between these indicators, we integrated the 298 significant indicators from different studies. After merging with a strict criterion (i.e., only indicators with similar meanings were merged), 230 indicators remained. Table 2 showed examples

Table 1
Studies included in the indicator analysis.

Index	Author, year	Participant number	Participant type	Index of mental health	Scale used in study	Number of proposed indicators	Number of significant indicators	Statistical test
1	Xie & Ye, 1994	220	Schizophrenia patients	Schizophrenia	Diagnosis	25	24	T test
2	Chen & Shen, 2004	290	Typical participants	Somatization	SCL-90	11	7	Regression
3	Chen & Xu, 2008	285	Typical participants	Depression	SCL-90	12	8	Regression
4	Qiu & Wu, 2010	35	Depressed patients	Depression	MMPI	NA	21	Correlation analysis
5	Chen et al., 2011	290	Typical participants	Anxiety	SCL-90	11	5	Regression
6	Zhu et al., 2011	706	Typical participants and participants with earthquake experience	PTSD, Anxiety	PCL-C, GHQ-28	176	39	ANOVA
7	Yan et al., 2013	1044	Typical participants	Suicide	Suicidal tendency scale (self –made)	13	13	Chi-square test
8	Li et al., 2014	105	HFA patients and typical participants	HFA	Raven, diagnosis	275	47	Chi-square test
9	Yan et al., 2014	540	Typical participants	Depression	SDS	15	13	Chi-square test
10	Deng, 2014	64	Schizophrenia patients and typical participants	Schizophrenia	BRPS	142	38	Chi-square test
11	(Zhao, Wang, et al., 2015)	296	Typical participants	Somatization	CSI	300	12	Chi-square test
12	Zhao, Peng, and Cheng, 2015	216	Depressed patients	Depression	HDRS	390	14	Correlation analysis
13	Ning et al., 2015	148	Depressed patients	Depression	CDI	62	31	Chi-square test
14	Tao et al., 2015	683	Typical participants	DPD	PDQ-4	15	7	Chi-square test
15	Sheng et al., 2019	167	Cancer patients	Anxiety	SAS	26	6	Chi-square test
16	Yang et al., 2019	167	Cancer patients	Depression	SDS	23	9	Chi-square test
17	Xiang et al., 2020	392	Typical participants	Depression	Achenbach CBCL	12	4	Chi-square test

Note. Index of mental health: PTSD = Post-traumatic stress disorder; HFA = High-functioning autism; DPD = Dependent personality disorder. Scales used in the study: SCL-90: Symptom Checklist-90; MMPI: Minnesota Multiphasic Personality Inventory; PCL-C: PTSD Check List–Civilian Version; GHQ-28: General Health Questionnaire-28; SDS: Self-Rating Depression Scale; CSI: Children's Somatization Inventory; HDRS: Hamilton Depression Rating Scale; CDI: Children's Depression Inventory; PDQ-4: Personality Diagnostic Questionnaire-Version 4; SAS: Self-Rating Anxiety Scale; Achenbach CBCL: Achenbach Child Behavior Checklist; BRPS: Brief Psychiatric Rating Scale. The 10th article is a master's thesis. Others are peer-review papers.

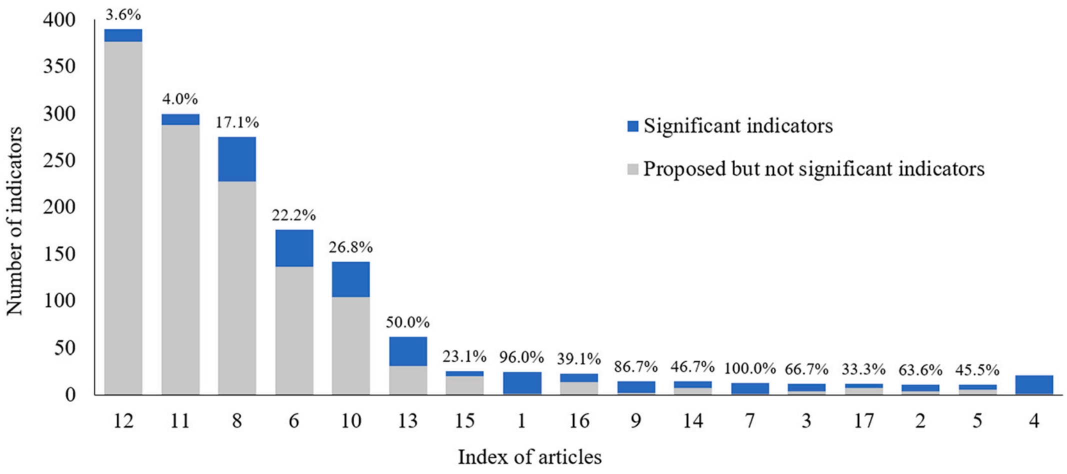

Fig. 4. Number and proportion of significant indicators.
Note. In this histogram, the light gray columns represent the number of indicators proposed but not significant by each article, and the blue (dark gray) columns represent the number of statistically significant indicators. Above each column is the percentage of significant indicators. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of the merged indicators. The indicators were divided into 5 categories (i.e., whole picture, house, tree, person, decoration), each of which can be subdivided into multiple sub-dimensions. The categories and sub-dimensions to which the indicators belonged were listed in Table 2. The related mental health issues and the number of articles supporting the relations were also included in Table 2.

Table 2
Examples of merged indicators.

Category	Sub-dimension	Indicator	Index of mental health	Number of supporting article
Whole picture	Thematic	Lack of drawing theme	Schizophrenia, HFA	2
Whole picture	Detail	Lack of details	Suicide, Depression	4
Whole picture	Size	Small drawing area	Schizophrenia, HFA, Depression	4
…	…	…	…	
House	Size	Small house	HFA, Depression, Schizophrenia, PTSD	4
House	Roof	Finely portray the roof	Anxiety	1
House	Door & window	Small windows and doors	Depression	1
…	…	…	…	
Tree	Trunk	Scars, stains or holes in the trunk	Somatization, DPD, PTSD, Depression, Anxiety	4
Tree	State	Withered	Suicide, Depression	2
Tree	Trunk	Paralleled trunk	PTSD, Depression	2
…	…	…	…	
Person	Body	Blank body	Depression, Somatization	5
Person	Hair	Chaotic hair	Schizophrenia, Depression, Anxiety	3
Person	Hand	Hands behind	Depression, Anxiety	3
…	…	…	…	
Decoration	Decoration	No sun	Suicide, Schizophrenia	2
Decoration	Decoration	Moon	Suicide, Depression	2
Decoration	Decoration	Fallen fruit	Depression	1
…	…	…	…	

Note. Decoration: elements other than house, tree, and human in the picture.

In the process of merging indicators, we found some contradictory indicators. For example, indicators of mouths open (Chen & Xu, 2008), not open (Yan et al., 2014), and no mouth (Ning et al., 2015) all showed significant associations with depression. With the strict criterion, these contradictory indicators were listed separately. Therefore, we also merged the indicators with a loose criterion, and the three indicators regarding mouth were integrated into one “mouth” indicator. With the loose criterion, 145 indicators remained.

After integrating all the significant indicators with two criteria, the number of articles supporting each indicator was obtained. Among the 230 indicators merged with the strict criterion (Fig. 5a), 183 indicators (79.6 %) were supported by only one paper, and only 10 indicators were supported by >2 papers. Even for the indicators merged with the loose criterion (Fig. 5b), 76 indicators (52.4 %) were supported by only one article, and only 16 indicators (11 %) were supported by 4 or more articles. The most consistent indicator was supported by 6 articles, less than half of the 17 articles.

In short, the small number of significant indicators in each study and the low overlapping of indicators across studies altogether suggested that the relationship between HTP indicators and mental health was unstable. However, it was possible that the existing indicators proposed in the indicator system were not the critical features related to mental health. Next, we used the DNNs to extract the features that might have been missed in the previous studies from the entire HTP drawings to explore if there was a relationship between mental health and HTP features in a higher-dimensional space.

3.3. Training DNNs to predict depression from HTP drawings

We collected the depression scores and HTP drawings of 4196 Chinese children and adolescents. The overall depression score was 42.5 ± 8.16. Divided by gender, the depression scores of males and females are 41.6 ± 7.73 and 43.4 ± 8.50 respectively. Divided by period of schools, the depression scores of primary school children, middle school adolescences and high school adolescences are 39.1 ± 8.29, 43.4 ± 8.19, 42.4 ± 7.75.

We first trained the pre-trained AlexNet to classify the HTP drawings of depression and typical individuals labeled with bisection criterion. The result was shown in Table 3. The test accuracy was 67.38 %, and the precision was 1, but the recall was only 0.007, so the F1 score was only 0.0139. In other words, the AlexNet classified all the test images as typical, except for one as depression. This might be due to the imbalanced sample size used in training ($N_{\text{depression}} = 1380$, $N_{\text{typical}} = 2809$).

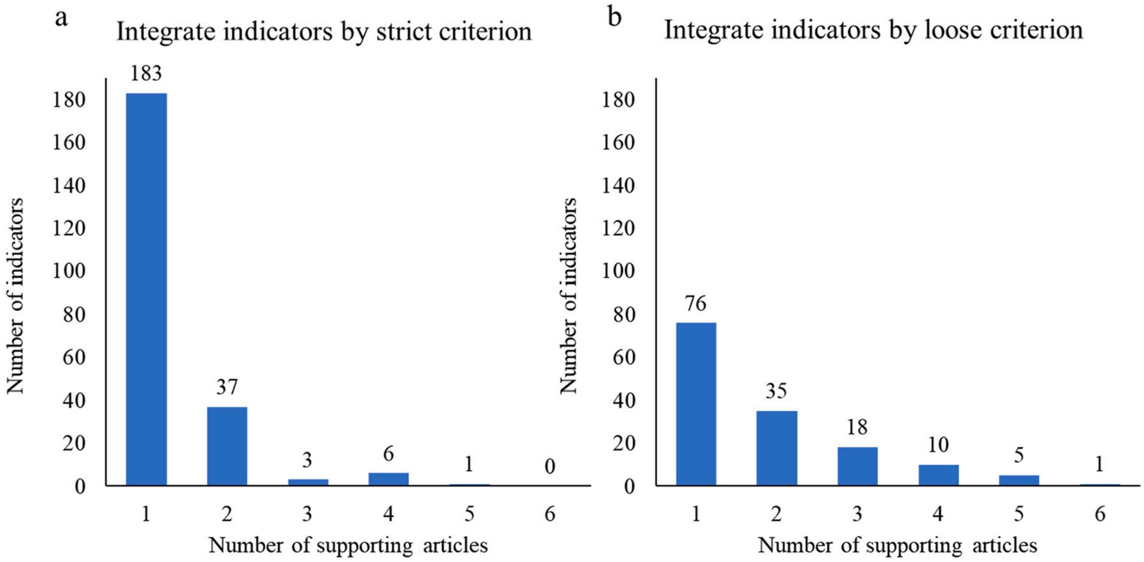

Fig. 5. The number of supporting articles for the indicators integrated with the strict criterion (a) and the loose criterion (b).

Table 3
Test results of AlexNet and InceptionV3 in classifying depression.

DNN architecture	Pre-training sample	Depression criterion	Test accuracy	Precision	Recall	F1 score
AlexNet	ImageNet	Bisection	67.38 %	1.0000	0.0070	0.0139
AlexNet	ImageNet	Trisection	58.82 %	0.6184	0.3821	0.4724
InceptionV3	ImageNet	Trisection	48.24 %	0.4824	1.0000	0.6508

When the DNN couldn't distinguish the images of depression and typical individuals, the best strategy was to classify all images into the larger category (i.e., typical), thus achieving optimal accuracy. In addition, the differences between the images of depression and typical individuals near the category boundary may be blurred when using bisection criterion.

In order to balance the sample size and increase the differences between depression and typical samples, we adopted the trisection criterion, in which only the two ends of the trisection sample were used. Using the samples labeled with trisection criterion, the test accuracy of the AlexNet was 58.82 % (Table 3, second row), lower than the bisection criterion result, but the F1 score was 0.4724, higher than the bisection criterion result. The precision and recall values (precision = 0.6184, recall = 0.3821) indicated that the AlexNet did not classify all the images as one category again. Nevertheless, the accuracy was only slightly higher than the chance level, which was far from accurate.

Next, we used another DNN, the InceptionV3, to classify depression and non-depression samples. Compared with AlexNet, InceptionV3 has a more complex architecture and achieves a better pre-training performance. As shown in Table 3, although the F1 score of the InceptionV3 was a little higher than the AlexNet, the test accuracy was only 48.24 %, even below the chance level.

To sum up, the DNNs pre-trained with ImageNet failed to learn the connection between HTP drawing and depression, even after considering the effects of imbalanced sample size, DNN architecture, and pre-training performance. However, the poor performance of the DNNs may be due to the big differences between pre-training images and the HTP images. The pre-training images were colorful natural images in ImageNet (Fig. 3b, left), which were quite different from the sketch HTP images (Fig. 3a). Previous study has shown that AlexNet used more texture than shape information when classifying images (Baker et al., 2018; Geirhos et al., 2019), while the sketch images lack texture information. Next, we changed the pre-training dataset from natural images to sketch images to test if the accuracy of depression classification would be increased.

3.4. *Training DNNs pre-trained with sketch images to predict depression*

To increase the similarity between the pre-training and training datasets, we changed the pre-training dataset from colorful natural images in ImageNet to black and white sketch images in Sketch dataset and our HTP drawings. We also slightly changed the DNN architecture by using the Sketch-A-Net for better classification of sketch objects. After pre-training, the Sketch-A-Net completed the object classification task of 293 categories with a validation accuracy of 78.36 %. Particularly, the Sketch-A-Net successfully classified the House, Tree, and Person images cropped from HTP drawings in the 293-category classification task, with an overall test accuracy of 94.92 % (96.19 % for House, 94.56 % for Tree, and 94.10 % for Person). A success in the object classification task indicated that the Sketch-A-Net was able to extract critical features of the house, tree, and person images.

Finally, we tested whether the Sketch-A-Net could use these features to predict depression. The results were shown in Table 4. The accuracies of the whole HTP, and cropped House, Tree, and Person images were only 55.69 %, 53.96 %, 50.17 %, and 56.62 % respectively. The precisions, recalls, and F1 scores were all around 0.5. These results didn't outperform the AlexNet significantly. This result showed that even though the Sketch-A-Net learned to extract critical features of the house, tree, and person images for object classification, it failed to classify the drawings of depressive individuals from typical individuals. And this failure was not due to the quality of our samples, since the same samples were successfully used by the Sketch-A-Net in object classification task.

4. Discussion

By combining meta-analysis and empirical research with DNNs, our study provided new empirical evidence questioning the validity of the HTP test in predicting mental health problems. First, we summarized previous studies on the relationship between HTP and mental health and found that there was no reliable correlation between the existing HTP indicators and mental health problems. Next, to examine the validity of the HTP test with a more objective approach, we used the DNNs to extract potential critical features from the entire HTP drawings. We found that although the DNNs successfully learned to extract critical features of houses, trees, and persons in HTP drawings for object classification, it failed to classify the drawings of depressive individuals from those of non-depressive individuals. Therefore, our results cast doubts on the validity of the HTP in diagnosing mental health problems.

In fact, the validity of the projective drawing test to predict mental health has been controversial since it was put forward (Lilienfeld et al., 2000). Since the indicator system was developed for the projective drawing test (Buck, 1948), many studies have tried to prove its validity with quantitative analyses such as *t*-test, correlation, or regression (Garb et al., 2002; Xiang et al., 2020; Xie & Ye, 1994; Yan et al., 2013, 2014). Although HTP and the indicator system have been commonly used in psychological counseling, there is no systematic investigation about whether the projective drawing tests can predict mental health. Our study systematically summarized previous studies on the relationship between HTP indicators and mental health. Our analysis revealed small number of significant indicators in each study and low consistency across studies about the relationship between HTP indicators and mental health, indicating that the HTP indicator system is not reliable. Consistent with our results, a previous meta-analysis study also concluded that the projective drawing tests cannot predict physical or sexual abuse (Allen & Tussey, 2012).

In addition to the summary of previous findings, our study also for

Table 4
Test results of Sketch-A-Net in classifying depression.

DNN architecture	Pre-training sample	Depression criterion	Picture	Test accuracy	Precision	Recall	F1 score
Sketch-A-Net	Sketch dataset+ Our sample	Trisection	HTP	55.69 %	0.5417	0.5285	0.5350
			House	53.96 %	0.5238	0.4925	0.5077
			Tree	50.17 %	0.4895	0.5000	0.4947
			Person	56.62 %	0.5479	0.5517	0.5498

Note. HTP = whole House-Tree-Person drawings.

the first time used DNNs to extract implicit features from entire HTP drawings that might have been missed in previous studies. The low accuracy of predicting mental health problems from HTP drawings even for experts in previous studies (Levenberg, 1975) may be due to the subjectivity of human expert judgement and coding procedure. In contrast, our study extracted HTP features objectively using DNNs, which has been successful to extract object features and achieve object classification task (Yu et al., 2017). From a data-driven perspective, DNNs enable us to explore the possible connection between features of HTP drawings and mental health more objectively and comprehensively. Even though the DNN could extract critical features to classify houses, trees and persons, it was still unable to learn the connection between HTP drawing and depression, which suggested that HTP cannot be used to predict mental health reliably. Combined with the meta-analysis study, these results cast doubts on the validity of the HTP in diagnosing mental health problems.

There is often a Q&A session between the consultant and the client about the drawing after test, which was not implemented in our study. It is possible that even if projective drawing tests are indeed effective, the relevant information is obtained from the one-to-one Q&A session. However, in the common application scenarios of projection drawing tests such as school group testing, there is no Q&A session. Therefore, the projective drawing test is inappropriate to be used to screen for mental problems on a large scale.

Notably, our sample only covered Chinese children and adolescents, which might limit the generalization of the present findings to broader population. Nevertheless, the culture difference in the features of children's HTP drawings is relatively small (Afolayan, 2015; Tomes & Fan, 2000). Future studies can further use deep learning to examine the validity of HTP in other cultures and races. Besides, we only studied the association between HTP drawing and depression with DNN. Future studies are invited to examine the associations between other projective drawing tests and other mental health problems using DNN as a tool.

5. Conclusion

Our study provides a novel empirical method to examine the validity of the projective drawing test using DNN, and casts doubts on the validity of HTP in diagnosing mental health problems.

CRediT authorship contribution statement

Yijing Lin: Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing—original draft, Writing—review and editing.

Nan Zhang: Data curation, Formal analysis, Investigation, Visualization, Writing—original draft, Writing—review and editing.

Yukun Qu: Methodology, Software.

Tian Li: Investigation, Resources.

Jia Liu: Conceptualization, Funding acquisition, Project administration, Supervision.

Yiying Song: Funding acquisition, Project administration, Supervision, Writing—review and editing.

Declaration of competing interest

None.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (31861143039, 31872786).

References

Afolayan, A. (2015). *Haitian children's House-Tree-Person drawings: global similarities and cultural differences*. Antioch University. Doctoral dissertation.

Allen, B., & Tussey, C. (2012). Can projective drawings detect if a child experienced sexual or physical abuse?: A systematic review of the controlled research. *Trauma, Violence, and Abuse, 13*(2), 97–111. https://doi.org/10.1177/1524838012440339

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS Computational Biology, 14*(12), Article e1006613. https://doi.org/10.1371/journal.pcbi.1006613

Blatt, S. J. (1975). The validity of projective techniques and their research and clinical contribution. *Journal of Personality Assessment, 39*(4), 327–343. https://doi.org/10.1207/s15327752jpa3904_1

Buck, J. N. (1948). The H-T-P test. *Journal of Clinical Psychology, 4*(2), 151–159. https://doi.org/10.1002/1097-4679(194804)4:2<151::AID-JCLP2270040203>3.0.CO;2-O

Catterall, M., & Ibbotson, P. (2000). Using projective techniques in education research. *British Educational Research Journal, 26*(2), 245–256. https://doi.org/10.1080/01411920050000971

Chen, K., & Shen, H.-Y. (2004). A research on the physical symptom in the projective drawing test. *Journal of Psychological Science, 27*(05), 1236–1238. https://doi.org/10.3969/j.issn.1671-6981.2004.05.060

Chen, K., Song, B., & Shen, H.-Y. (2011). Using the projective drawing test to evaluate the anxiety symptom. *Journal of Psychological Science, 34*(06), 1512–1515. https://doi.org/10.16719/j.cnki.1671-6981.2011.06.042

Chen, K., & Xu, G.-X. (2008). A research on the diagnosis of depression through the projective drawing test. *Psychological Science, 31*(03), 722–724. https://doi.org/10.3969/j.issn.1671-6981.2008.03.050

Tao, C., Pei, H., Wang, P., Xing, Y., Luo, J., & Xiang, J. (2015). On the diagnosis of teenagers' dependent personality disorder inclination-based on the projective drawing test of S-HTP. *Chinese Journal of Special Education, 176*(2), 59–64. https://doi.org/10.3969/j.issn.1007-3728.2015.02.010

Chen, X., Zhou, M., Gong, Z., Xu, W., Liu, X., Huang, T., Zhen, Z., & Liu, J. (2020). DNNBrain: A unifying toolbox for mapping deep neural networks and brains. *Frontiers in Computational Neuroscience, 14*. https://doi.org/10.3389/fncom.2020.580632

Ciregan, D., Meier, U., & Schmidhuber, J. (2012, June). Multi-column deep neural networks for image classification. In *2012 IEEE conference on computer vision and pattern recognition* (pp. 3642–3649). IEEE. https://doi.org/10.1109/CVPR.2012.6248110.

Deng, C.-Y. (2014). *The study of correlation between Schizophrenics SHTP test and BPRS*. Guangzhou University of Chinese Medicine. Master dissertation.

Eitz, M., Hays, J., & Alexa, M. (2012). How do humans sketch objects? *ACM Transactions on Graphics, 31*(4), 1–10. https://doi.org/10.1145/2185520.2185540

Garb, H. N., Lilienfeld, S. O., Wood, J. M., & Nezworski, M. T. (2002). Effective use of projective techniques in clinical practice: Let the data help with selection and interpretation. *Professional Psychology: Research and Practice, 33*(5), 454–463. https://doi.org/10.1037/0735-7028.33.5.454

Garb, H. N., Wood, J. M., & Teresa Nezworski, M. (2000). Projective techniques and the detection of child sexual abuse. *Child Maltreatment, 5*(2), 161–168. https://doi.org/10.1177/1077559500005002007

Geirhos, R., Michaelis, C., Wichmann, F. A., Rubisch, P., Bethge, M., & Brendel, W. (2019). Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *7th International Conference on Learning Representations, ICLR 2019*.

Goodenough, F. L. (1928). Studies in the psychology of children's drawings. *Psychological Bulletin, 25*(5), 272–283. https://doi.org/10.1037/h0071049

Goodenough, F. L., & Harris, D. B. (1950). Studies in the psychology of children's drawings: II 1928–1949. *Psychological Bulletin, 47*(5), 369–433. https://doi.org/10.1037/h0058368

Kovacs, M. (1985). The children's depression, inventory (CDI). *Psychopharmacology Bulletin, 21*(4), 995–998.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (p. 25).

Levenberg, S. B. (1975). Professional training, psychodiagnostic skill, and kinetic family drawings. *Journal of Personality Assessment, 39*(4), 389–393. https://doi.org/10.1207/s15327752jpa3904_11

Li, X., Cao, B.-D., Yang, W., Qi, J.-H., Liu, J., & Wang, Y.-F. (2014). Characteristic of the synthetic house-tree-person test in children with high-functioning autism. *Chinese Mental Health Journal, 28*(4), 260–266. https://doi.org/10.3969/j.issn.1000-6729.2014.04.005

Lilienfeld, S. O., Wood, J. M., & Garb, H. N. (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest, 1*(2), 27–66. https://doi.org/10.1111/1529-1006.002

Little, R. J. A., & Rubin, D. B. (2014). Statistical analysis with missing data.. In *Statistical analysis with missing data*. https://doi.org/10.1002/9781119013563

Joiner, T. E., Jr., Schmidt, K. L., & Barnett, J. (1996). Size, detail, and line heaviness in children's drawings as correlates of emotional distress:(More) negative evidence. *Journal of Personality Assessment, 67*(1), 127–141. https://doi.org/10.1207/s15327752jpa6701_10

Ning, S.-Y., Zheng, L., Li, X., & Hui, W.-J. (2015). Application of house-tree-person test in evaluating adolescent depression. *Chinese Journal of Clinical Research, 28*(03), 305–307. https://doi.org/10.13429/j.cnki.cjcr.2015.03.011

Qiu, H., & Wu, D. (2010). Correlation study on MMPI and HTP drawing characteristics of depression patients. *China Journal of Health Psychology, 18*(11), 1341–1344. https://doi.org/10.13342/j.cnki.cjhp.2010.11.014

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision, 115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Sangkloy, P., Burnell, N., Ham, C., & Hays, J. (2016). The sketchy database. *ACM Transactions on Graphics, 35*(4), 1–12. https://doi.org/10.1145/2897824.2925954

Sheng, L., Yang, G., Pan, Q., Xia, C., & Zhao, L. (2019). Synthetic house-tree-person drawing test: A new method for screening anxiety in cancer patients. *Journal of Oncology, 2019*, Article 5062394. https://doi.org/10.1155/2019/5062394

Sladojevic, S., Arsenovic, M., Anderla, A., Culibrk, D., & Stefanovic, D. (2016). Deep neural networks based recognition of plant diseases by leaf image classification. *Computational Intelligence and Neuroscience, 2016*, Article 3289801. https://doi.org/10.1155/2016/3289801. Master dissertation.

Sun, Y., Liang, D., Wang, X., & Tang, X. (2015). *Deepid3: Face recognition with very deep neural networks*. https://doi.org/10.48550/arXiv.1502.00873. arXiv preprint arXiv:1502.00873.

Szegedy, C., Toshev, A., & Erhan, D. (2013). Deep neural networks for object detection. In *Advances in neural information processing systems* (p. 26).

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-December* (pp. 2818–2826). https://doi.org/10.1109/CVPR.2016.308

Tomes, R. E., & Fan, C. (2000). A cross-cultural comparison of draw-a-person, draw-a-house-tree, and piagetian cognitive tasks for Chinese and American children. *Journal of Early Childhood Teacher Education, 21*(2), 295–301. https://doi.org/10.1080/0163638000210224

Wildman, R. W., & Wildman, R. W., II. (1975). An investigation into the comparative validity of several diagnostic tests and test batteries. *Journal of Clinical Psychology, 31*(3), 455–458. https://doi.org/10.1002/1097-4679(197507)31:3<455::AID-JCLP2270310319>3.0.CO;2-1

World Health Organization. (2004). *Promoting mental health: Concepts, emerging evidence, practice: Summary report*. World Health Organization.

World Health Organization. (2017). *Depression and other common mental disorders: global health estimates (No. WHO/MSD/MER/2017.2)*. World Health Organization.

Wu, W.-F., Lu, Y.-B., Tan, F.-R., & Yao, S.-Q. (2010). Reliability and validity of the Chinese version ofchildren's depression inventory. *Chinese Mental Health Journal, 24*(10), 775–779. https://doi.org/10.3969/j.issn.1000-6729.2010.10.014

Xiang, J.-J., Liao, M.-S., & Zhu, M.-J. (2020). Assessment of junior elementary pupils' depression tendency via house-tree-person test. *China Journal of Health Psychology, 28*(7), 1057–1061. https://doi.org/10.13342/j.cnki.cjhp.2020.07.023

Xie, L.-Y., & Ye, X.-H. (1994). Primary application of synthetic house-tree-person technique in China: A comparison of schizophrenics and typical controls. *Chinese Mental Health Journal, 8*(06), 250–252,286.

Yan, H., Yang, Y., Wu, H.-S., & Chen, J.-D. (2013). Applied research of house-tree-person test in suicide investigation of middle school students. *Chinese Mental Health Journal, 27*(09), 650–654. https://doi.org/10.3969/j.issn.1000-6729.2013.09.002

Yan, H., Yu, H., & Chen, J. (2014). Application of the house-tree-person test in the depressive state investigation. *Chinese Journal of Clinical Psychology, 22*(5), 842–845. https://doi.org/10.16128/j.cnki.1005-3611.2014.05.065

Yang, G., Zhao, L., & Sheng, L. (2019). Association of synthetic house-tree-person drawing test and depression in cancer patients. *BioMed Research International, 2019*, Article 1478634. https://doi.org/10.1155/2019/1478634

Yu, Q., Yang, Y., Liu, F., Song, Y. Z., Xiang, T., & Hospedales, T. M. (2017). Sketch-a-Net: A deep neural network that beats humans. *International Journal of Computer Vision, 122*(3), 411–425. https://doi.org/10.1007/s11263-016-0932-3

Zhao, W.-J., Peng, Y., & Cheng, S.-Y. (2015). Correlation analysis between adolescent depression and House-Tree-Person test. *Modern Communication, 410*(6), 124–125. https://doi.org/10.3969/j.issn.1009-5349.2015.06.089

Zhao, Y., Wang, Q.-Y., Xiang, J.-J., & Wang, Q. (2015). Drawing characteristics of somatization tendency children in house-tree-person test. *Chinese Mental Health Journal, 29*(2), 115–120. https://doi.org/10.3969/j.issn.1000-6729.2015.02.007

Zhu, H.-L., Xiang, J.-J., Chen, W.-J., Shen, H.-Y., & Gao, L. (2011). House-tree-person painting characteristics of adolescents with post-traumatic stress disorder in Sichuan earthquake area. *Journal of Educational D, 06*, 39–42. https://doi.org/10.16215/j.cnki.cn44-1371/g4.2011.06.002