arXiv:2106.15323v2 [cs.CV] 1 Jul 2021

FACE IDENTIFICATION PROFICIENCY TEST DESIGNED USING ITEM RESPONSE THEORY

A PREPRINT

Géraldine Jeckeln
The University of Texas at Dallas

Ying Hu
The University of Texas at Dallas

Jacqueline G. Cavazos
The University of Texas at Dallas

Amy N. Yates
National Institute of Standards and Technology

Carina A. Hahn
National Institute of Standards and Technology

Larry Tang
University of Central Florida

P. Jonathon Phillips
National Institute of Standards and Technology

Alice J. O'Toole
The University of Texas at Dallas

July 2, 2021

ABSTRACT

Measures of face identification proficiency are essential to ensure accurate and consistent performance by professional forensic face examiners and others who perform face identification tasks in applied scenarios. Current proficiency tests rely on static sets of stimulus items, and so, cannot be administered validly to the same individual multiple times. To create a proficiency test, a large number of items of "known" difficulty must be assembled. Multiple tests of equal difficulty can be constructed then using subsets of items. Here, we introduce a proficiency test, the Triad Identity Matching (TIM) test, based on stimulus difficulty measures based on Item Response Theory (IRT). Participants view face-image "triads" (N=225) (two images of one identity and one image of a different identity) and select the different identity. In Experiment 1, university students (N=197) showed wide-ranging accuracy on the TIM test. Furthermore, IRT modeling demonstrated that the TIM test produces items of various difficulty levels. In Experiment 2, IRT-based item difficulty measures were used to partition the TIM test into three equally "easy" and three equally "difficult" subsets. Simulation results indicated that the full set, as well as curated subsets, of the TIM items yielded reliable estimates of subject ability. In summary, the TIM test can provide a starting point for developing a framework that is flexible, calibrated, and adaptive to measure proficiency across various ability levels (e.g., professionals or populations with face processing deficits).

***K*eywords** face identification · item response theory · 3-AFC

1 Introduction

Face identification is a critical element of law enforcement and criminal justice in the United States and abroad. This important task is carried out by professional forensic face examiners. Therefore, it is incumbent on the justice system to ensure that professional face examiners exhibit optimal performance on face identification tasks and that their proficiency is sustained over time. However, limited work has been done to develop tests that enable the assessment of proficiency (a person's internal ability for face identification which can be inferred from their accuracy on a test) across time. To address this, we propose a novel framework that enables face-identification testing across time and across individual ability levels.

Reliable measures of proficiency at different points in time are desirable for several reasons. For example, individuals in professions that require them to make face identification decisions (e.g., forensic face examiners in law enforcement) will often participate in training courses designed to improve their accuracy. Proficiency measures gathered over time can gauge the effectiveness of these training programs (e.g., scores before vs. after training). Also, these measures can be used to assess the effects of experience and age on proficiency to assure sustained proficiency over time. To measure accuracy at different time points, we need to conceptualize a proficiency test in terms of multiple equally-difficult tests composed of items sampled from a large pool of stimuli. For this reason, in the present study, we developed a face identification test that can provide reliable proficiency estimates across time and across individual ability levels.

We were motivated by the special problems of forensic examiners due to their role in social justice and safety. However, to develop the framework, we studied the skills of people from the general population (e.g., undergraduates), with the goal of developing a test that can be applied across a wide range of subject abilities. This variability would make the test applicable to people of high ability (e.g., super-recognizers [1, 2, 3]), as well as to clinical populations in which individuals exhibit atypical face processing skills, for example, autism spectrum disorders [4] and schizophrenia [5].

A proficiency test should have two properties. First, it should support the creation of multiple sub-tests of equal difficulty. Sub-tests are needed, because a single test cannot be taken more than once. Repeated exposure to the same faces can inflate identification accuracy via familiarity effects [6]. This repeated exposure is highly problematic when evaluating training programs, because it can appear as if a test takers' general skills have improved, when increased accuracy is due to familiarity with the stimuli. Second, a proficiency test should be calibrated. That is, it should contain stimulus items of "known" difficulty that can be sampled (without replacement) and stratified into graded difficulty levels. This supports the creation of tests that stratify subjects into graded levels of ability.

Measuring item difficulty in existing proficiency tests [e.g., Glasgow Face Matching Test [7], Kent Face Matching Test [8]] is complicated by the problem of the response bias of test takers (i.e., a user's internal tendency to select one response category over another). Before we expand on this issue, we digress to provide background on how an identification task is performed.

Identity-matching is the most commonly used task for assessing the proficiency of professional face examiners. In this task, examiners compare two face images (e.g., security camera image vs. mugshot) and must indicate whether the images show the same person or different people. The tests require either a binary response ("same" or "different" person) or a response rating (e.g., -2: Sure they are the different; +2: Sure they are the same). In both cases, response bias can be exhibited via the observer's tendency to select one response option over another. Bias can be due to the observer's internal decision criterion [9, 10], differential use of the Likert-type scale [11, 12], or to situational factors such as the perceived cost of certain types of incorrect decisions (identify or fail to identify).

To illustrate how response bias complicates item difficulty measures, let's consider a face identification task with binary response options (i.e., "same" or "different" identity). When uncertain about an identification decision, an observer with a conservative response bias will exhibit a greater tendency to respond "different identity" in comparison to an observer with a liberal response bias. Considered from the perspective of item difficulty, a conservative response bias results in greater accuracy for different-identity pairs than for same-identity pairs. Thus, different-identity pairs would appear (incorrectly) to be easier than same-identity pairs [11]. The opposite is true for liberal observers. Alternately, when a response rating is made on a Likert scale, item difficulty would be gauged by relative "confidence" for same- versus different-identity items. For instance, a same-identity pair that receives a response of +1 (Think they are the same) would be assumed to be more difficult than a same-identity pair that receives a response of +2 (Sure they are the same).

Consequently, for identity-matching tasks, the most common type of face identification test, observer criterion and item difficulty measures are co-dependent. This is true regardless of whether participants make binary or rated responses. This is a serious problem for cases in which groups of participants are compared. Specifically, when there are group-based differences in response bias (e.g., students, forensic examiners), item difficulty comparisons across groups are not valid [11].

Previous studies do, in fact, show group-based differences in the use of response scales. Forensic examiners, compared to untrained undergraduates, concentrate their responses in the middle of the scale (less certain), thereby avoiding "high confidence" responses at extreme ends of the scale [11, 12]. Examiners may adopt this strategy to avoid the repercussions of high-confidence misidentifications in forensic face settings. This compromises the validity of item difficulty measures applied across groups of individuals (e.g., forensic examiners, students). Again, from the perspective of item difficulty, many items would be found to be more difficult for examiners (higher-ability individuals) than for students (lower-ability individuals) [11, 13], when instead the item difficulty measure is driven by the differential use of the scale by the two populations. One theoretical approach to measuring item difficulty directly is to use Item Response Theory (IRT [14]). Before presenting the main part of the study, we first introduce IRT for assessing item difficulty and subject ability.

1.1 Item Response Theory

In recent research on face perception [15, 16, 17, 18], IRT has been proposed as a method for test evaluation and ability assessment. IRT is a psychometric theory used to model the association between face-identification decisions (subject responses to items) and face identification ability (see Section 7). IRT offers a promising route for testing identification ability, because it provides estimates of ability based on the properties of items. Also, IRT has several features that are critical for developing a face identification test. We consider these in turn.

First, IRT provides measures of subject ability and item difficulty that occupy the same scale[1] and can be compared directly to one another [19]. This property is particularly valuable for building proficiency assessment tools that are intended to capture specific levels of face identification skill. As illustrated in Fig. 1, this important feature enables the user to infer each subjects' probability of responding to a specific item correctly, given their respective position on the ability scale.

Second, IRT provides item difficulty measures that are independent of the subject sample [19]. Concomitantly, IRT provides ability measures that are independent of the item sample and can be generalized to the subjects' true skill level [19].

Third, IRT provides precision measurements at the individual subject ability level. This feature of IRT enables evaluation of the test for assessing people of different ability via the Test Information Function [19, 15]. The peak of this function corresponds to the subject ability level that is best suited for evaluation with the test. For example, IRT can assess the efficiency of existing assessment tools for diagnosing individuals with impaired face recognition [16].

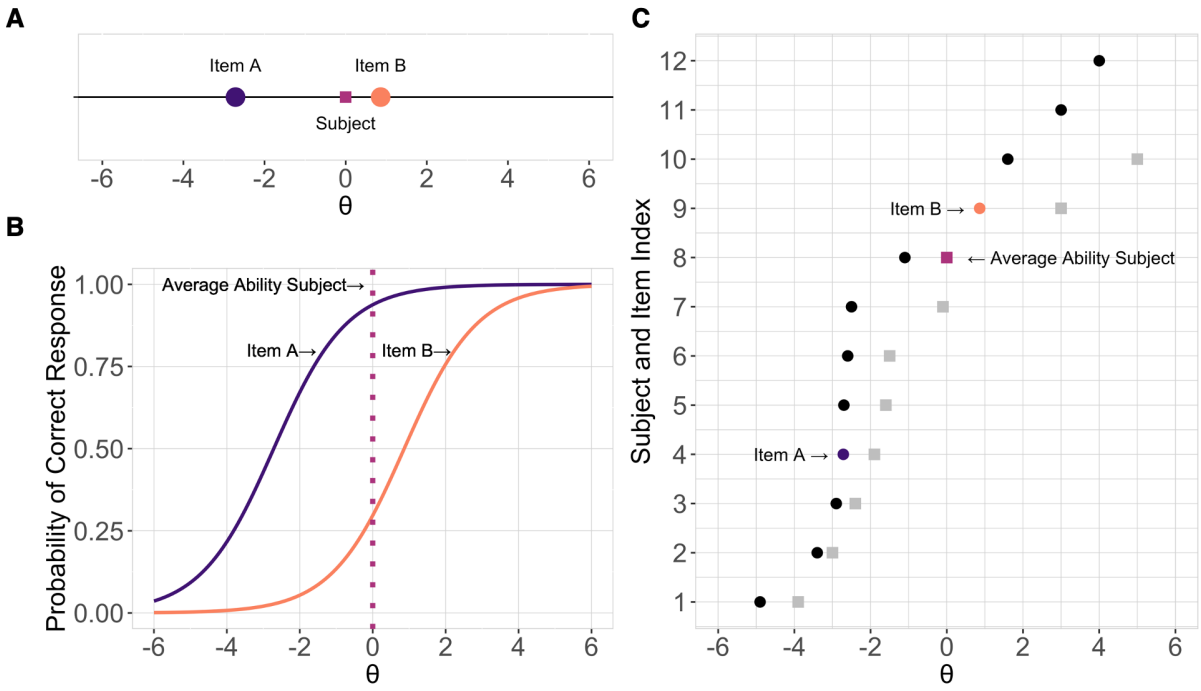

Figure 1: A) Example of IRT Subject and Item Scale. Item difficulty (β) and subject ability (θ) occupy the same latent scale ranging from low (easy item, poor ability) to high (difficult item, high ability). By convention, this shared scale is labeled (θ). In the model used in this paper, average ability is defined as 0. One exemplar subject (magenta square) is used to represent average ability. If a subject's estimated ability (square) is greater than the estimated difficulty of an item (circles), the subject has an above-chance probability to answer the given item correctly. This can be seen more clearly when the probability of a correct response is plotted as the function of subject ability. B) For example, a subject with average ability (dotted line, θ=0) has an above-chance probability (.94) of endorsing a correct response to Item A (B, β = -2.72) and below-chance probability (.30) of endorsing a correct response to Item B (C, β = 0.86). C) Exemplar items (circles) and subjects (squares) plotted along the ability and difficulty scale. The index consists of items and subjects ranked by difficulty and ability, respectively.

[1]Following conventions used in the IRT literature, we refer to this as the "Theta scale," henceforth labeled θ.

In previous studies, IRT was used to evaluate the quality of face recognition tests [16], to isolate face recognition ability from other abilities [15], and to assess item bias towards certain demographic groups [17]. These previous studies assess *memory* for faces. IRT has not yet been applied to face identification tasks that are based on *perception* (without memory requirements) (See Section 8.1, Table 3). These perceptual abilities are tapped in forensic identification (e.g., identity matching). Here, we use IRT to analyze the results of face identification tests that rely on perceptual abilities.

Specifically, our goal for this study was to develop a face identification test that enables testing across time and across ability levels. To achieve this goal, first we propose a three alternative forced choice (3-AFC) face identification test (the Triad Identity Matching [TIM] test). A 3-AFC paradigm can be used to construct calibrated subsets of items, because it allows for item difficulty estimates that circumvent the response bias issues of existing tests. Therefore, the second step was to use IRT to measure the psychometric properties of the TIM test. Item difficulty scores extracted from the IRT modelling were used to create subsets of stimulus items that can be partitioned into equal difficulty levels or stratified into various graded difficulty levels ("easy" or "difficult"). To date, no study has considered the usefulness of IRT for evaluating the psychometric properties of face identity matching tasks.

The following text is organized into four sections. In Section 2, we describe the TIM test construction. In Section 3 (Experiment 1), we provide data from university students on the TIM test and evaluate the test using IRT, along with traditional measures of item difficulty and subject accuracy (proportion correct). In Section 4 (Experiment 2), we demonstrate how IRT can be used to guide the construction of equally difficult sub-tests, and provide comparisons between ability estimates computed from the subsets of items and the full test. In Section 5 (Discussion), we conclude with the contributions and limitations of this work.

2 TIM Test Construction

We created the TIM test, a 3-AFC test, consisting of image triads: two same-identity images and one different-identity image. Participants determine which of the images depicts the different-identity ("odd-one-out") (Figure 2). A total of 225 triads were created using 675 images sampled from the *Good, Bad, and Ugly Face Challenge Dataset* [20]. Images were taken in frontal view and varied in illumination, expression, and subject appearance (e.g., accessories and hair).

To avoid ceiling effects, triads were constructed to minimize the similarity of images that showed the same identities. The different-identity image in the triad was chosen to be as similar as possible to one of the same-identity images (Figure 2). Images for triads were selected using data from VGG-Face [21]. We selected this algorithm, because it proved comparable in ability to students in a face identity matching test (cf. A2015 [12]). Here, we used the top-level face descriptors from the algorithm to compute similarity scores between images. In what follows, for any identity pair A and B, a triad includes two images of one identity (A_0 and A_i) and one image of a different identity (B_0).

Figure 2 shows a more detailed account of the following step-wise process for triad construction.

1. Image pairs were divided into same- and different-identity pairs.
2. Similarity scores were used to rank pairs of different-identity images from the most to the least similar.
3. The different-identity image pair (A_i, B_j) with the largest similarity score was selected.
4. Similarity scores were also used to rank same-identity images (A_i, A_j) from the most to the least similar. The least similar image was chosen to complete the triad.

Therefore, each triad consisted of different-identity pairs that are similar and a same-identity pair that is dissimilar (A_0, A_i, and B_0). By design, the algorithm should perform poorly on the triads. We verified the difficulty of the TIM test for the algorithm, as follows. We treated VGG-face as a subject and simulated the 3-AFC face identification task. The odd-one-out decision was made using the algorithm-generated similarity scores. For each triad, the two images with the highest similarity were judged as the same identity and the remaining image was selected as the odd-one-out. The selection was compared to the ground truth and the proportion correct was calculated. The result showed systematically incorrect performance (proportion correct = .1378) for the algorithm, thereby supporting the use of similarity scores from VGG-face [21] to construct highly challenging triads.

3 Experiment 1 - Normative Performance on the TIM Test

In Experiment 1, we demonstrate that the TIM test can capture a large range of student performance and individual item accuracy. In addition, we show that using IRT-based parameters we can obtain item difficulty and subject ability measures on the TIM test. First, student performance was evaluated on the full set of TIM stimuli. Individual student baseline accuracy was calculated as the proportion of correct responses. Second, we employed IRT modeling (Rasch, 1960, [22]) to evaluate the psychometric properties of the test.

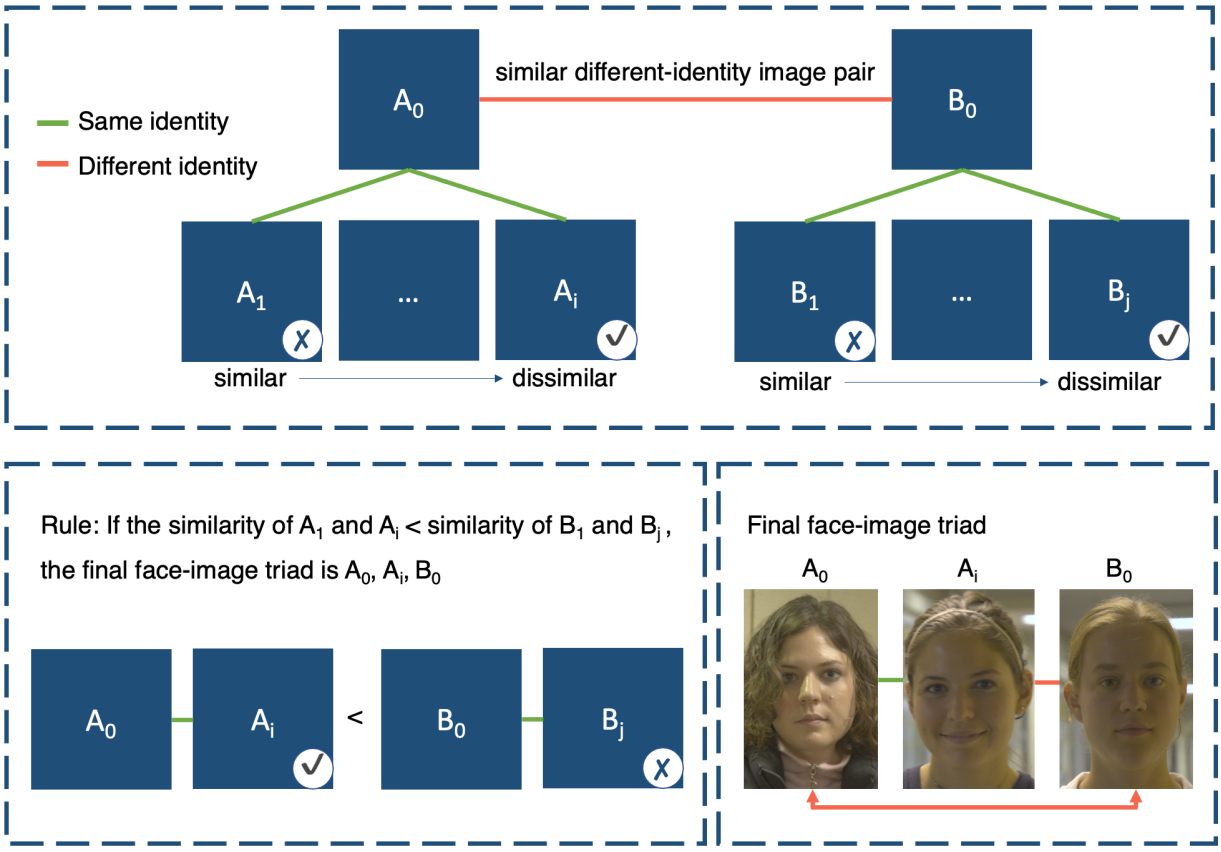

Figure 2: Stimuli screening paradigm. Image pairs were divided into same-identity and different-identity pairs. Then, different-identity pairs were ranked from similar to dissimilar using similarity scores obtained from a deep convolutional neural network (VGG-face, [21]). All different-identity pairs were demographically constrained ("yoked") so that only the same-race and same-gender pairs remained. Next, for each identity, only the different-identity pair with the largest similarity score was chosen. For each identity within a different-identity pair, a same-identity image with the lowest similarity was selected. The identity with the lowest same-identity pair similarity score was selected to be part of the triad. Therefore, each triad consisted of the lowest different-identity pair and the highest same-identity pair for that identity.

3.1 Methods

Participants. A total of 203 undergraduate students from The University of Texas at Dallas (UTD) participated in this study. Data collection took place during the Spring 2019 (77 participants) and (early) Spring 2020 (126 participants) semesters. Participants were recruited through The School of Behavioral and Brain Sciences online sign-up system and were compensated with research exposure credits. Participants were required to be at least 18 years of age and have normal- or corrected-to-normal vision. Datasets from two participants were removed, because they were incomplete, and datasets from four participants were removed for missing data. The analysis included 197 participants (140 female, 55 male, and 2 indicated "other"), ranging from age 18 to 36 (average age = 20.23). All aspects of the study were in accordance with the UTD Institutional Review Board protocol.

Procedure. For each participant, data collection took place in a single experimental session and included the full TIM test (225 items), followed by a demographic survey (via Qualtrics [23][2]). The experiment was programmed using

[2]Certain commercial equipment, instruments, or materials are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

PsychoPy v1.84.2 [24]. For each trial, a triad was presented for 3.5 seconds. Response time was not limited and no feedback was provided. Trial order and image position within a triad were randomized across participants.

3.2 Analysis and Results

Baseline Performance. Performance was measured as the proportion of items answered correctly. Chance performance was 0.33. Participant accuracy was well above chance (M= 0.69, SD= 0.11, Mdn= .70) and ranged from 0.37 to 0.89. Item accuracy (proportion of subjects who answered each item correctly) varied widely, ranging from .17 to .97 (M= .69, SD= .17, Mdn= .72).

IRT Modeling. We employed IRT modeling to evaluate the psychometric properties of the TIM test (see Figure 3). A one-parameter logistic model (Rasch model, [22]) was fit to the data employing Expectation Maximization (EM). All aspects of IRT modeling were conducted in R, using the mirt package v1.29 [25]. A scree test [26] was used to evaluate the dimensionality of the data and to ensure that the TIM test measured a single latent variable (face identification ability). Model fit was assessed using the root mean square error of approximation (RMSEA), Akaike information criterion (AIC), and Bayesian information criterion (BIC). A RMSEA of .6 and below is considered a good model fit.

Scree test results indicated unidimensional data. Results also indicated a good fit for the one-parameter logistic model (RMSEA = 0, AIC = 47195.18, BIC = 47937.19). Figure 3A shows fitting responses of 197 subjects on 225 triads. Triad difficulty spanned -3.81 to 1.67, and subject ability spanned -1.53 to 1.29. An efficient proficiency test should capture accurate estimates of proficiency for different ability groups. To simulate this, we show that IRT models built from groups with high (low) ability individuals can accurately measure the proficiency of groups with low (high) regardless of differences in abilities (See 8.2 and Figure 7).

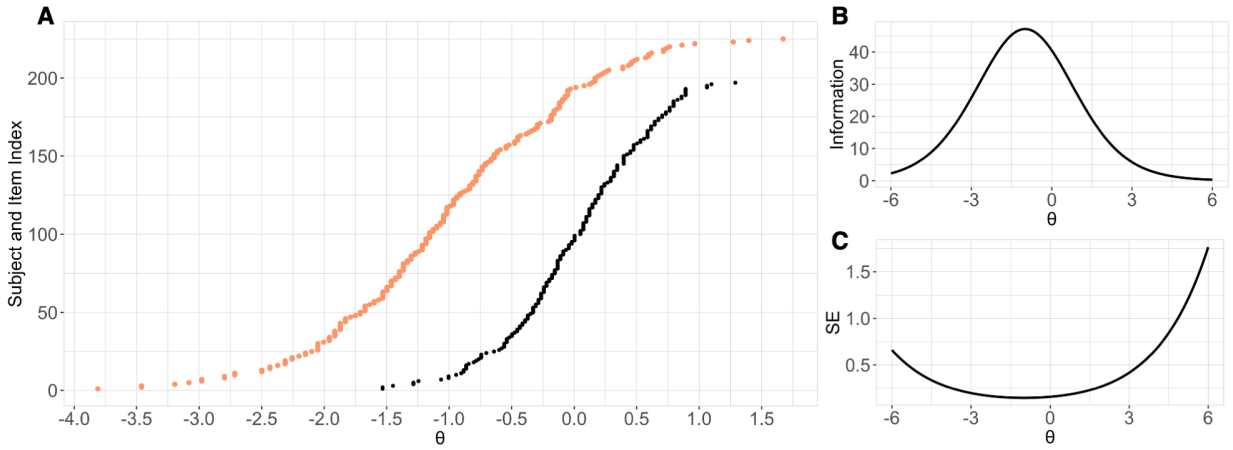

Figure 3: (A) Rasch model fit to 225 items and 197 subjects. Item difficulty (orange) and subject ability (black) estimates are plotted on the same scale. (B) The test information function is the reciprocal of the standard error of the estimated construct (θ) and is commonly used to indicate the degree of measurement precision for any ability score. The results suggests that the TIM test is most informative for ability scores ranging between low to average. (C) The standard error of the estimated construct (θ).

For both item difficulty and subject ability, linear regions overlapped. The peak of the Test Information Function demonstrated that the test provided the most reliable estimates of ability for subjects with low (θ = -1.5) to average (θ = 0) ability (Figure 3B). The Test Information Function is a standard measure in IRT. However, this information can also be conveyed using psychological standard measures such as standard error (SE) of the estimated construct (Figure 3C). The SE curve is the inverse of the Test Information Function, and demonstrates also that the lowest error estimates (lowest point of the curve) were for low-to-average ability individuals.[3]

[3]Baseline performance on an initially smaller sample size (N = 76), (M= 0.68, SD= 0.11, Mdn= 0.69), was comparable to our final sample size (N = 197). Item difficulty measures also were correlated across subject sample sizes [proportion correct: r = .97, p <.0001; β: r = .97, p <.0001].

3.3 Experiment 1 Discussion

The TIM test captured a large range of student performance and individual item accuracy. The test was particularly informative for subjects with low-to- average abilities. The capacity of the IRT to see both subject ability and item difficulty simultaneously offered assurance that the range of item difficulty spanned the subject ability range.

4 Experiment 2 - Creating Subsets of Customized Difficulty

In Experiment 2, we show that the TIM test can be partitioned into effective and equally challenging sub-tests. Equally-difficult sub-tests are crucial for recruitment and training purposes, particularly in applied scenarios (forensic facial examiners). Furthermore, item subsets of graded difficulty are needed for testing subject groups of different ability. Therefore, the main goal of Experiment 2 was to investigate the application of IRT to create subsets of items of particular difficulty and to test the reliability of subtests made from these items.

4.1 Methods

The TIM test items ($n = 225$) were partitioned into a total of six 36-item subsets. First, items were ranked from most easy to most difficult based on item difficulty measures obtained from the Rasch model trained on student data (Experiment 1). Next, the ranked items were median split into an easy and a difficult set. Finally, items from each easy and difficult set were randomly sampled (without replacement) to create three “Easy" and three “Difficult" subsets.

4.2 Results

Baseline Performance. All analyses were carried using the data collected in Experiment 1 (197 university students). Item difficulty measures were derived from the Rasch model trained on the university student data. Descriptive statistics for item difficulty are reported in Table 1 and Figure 4. As expected, item difficulty scores were greater for the three “Difficult” sub-tests and lower for the three “Easy” sub-tests.

Table 1: Describtive statistics for item difficulty (β) for the six subsets in Experiment 2.

Subset	*M*	*SD*	*Mdn*
Easy 1	-1.72	.47	-1.66
Easy 2	-1.72	.67	-1.48
Easy 3	-1.84	.64	-1.67
Difficult 1	-.30	.55	-.30
Difficult 2	-.23	.60	.41
Difficult 3	-.18	.55	-.17

Baseline performance was measured as the proportion of items answered correctly for each subset. Descriptive statistics are reported in Table 2 and plotted as a violin plot in Figure 5A. As expected, accuracy was higher for the “Easy” sub-tests than for the “Difficult” sub-tests. In addition, we compared proportion correct on the sub-tests against proportion correct on the full test. Pearson product-moment correlation results indicated a strong positive relationship between the full TIM item bank and each subset of items (r = .81 - .88)(see Figure 5B). Comparisons across all subsets showed a moderate positive relationship, ranging from (r = .58) to (r = .77). This range of results is as expected and is consistent for an IRT model that fits well.

Table 2: Descriptive statistics for subject accuracy (proportion correct) on the six subsets in Experiment 2.

Subset	*M*	*SD*	*Mdn*
Easy 1	.83	.11	.83
Easy 2	.82	.12	.83
Easy 3	.84	.11	.86
Difficult 1	57	.14	.58
Difficult 2	.55	.14	.56
Difficult 3	.54	.14	.53

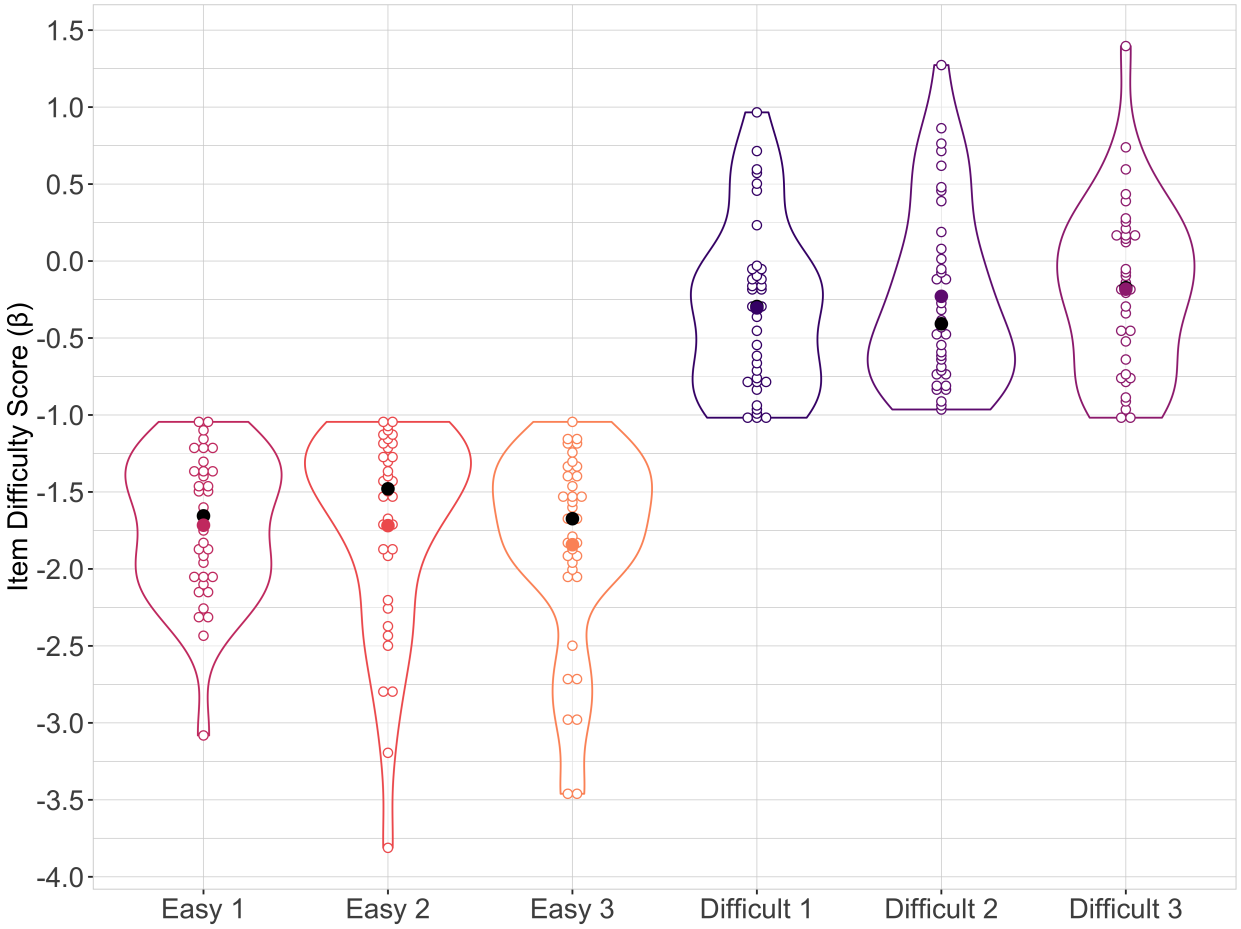

Figure 4: Violin plots of item difficulty on the six subsets. The empty circles represent the difficulty scores of individual items, the coloured dots represent the item subset mean, and the black dots represent the item subset median. Average item difficulty was lower for all three "Easy" subsets compared to all three "Difficult" subsets.

Ability Estimates. Here we demonstrate that the TIM test produces consistent estimates of subject ability with different test sizes. Specifically, we show that smaller subsets of items converge to give similar estimates of subject ability. We do this by examining the Rasch model trained on university students using the full set of TIM items (Experiment 1). Particularly, we tested whether the model can be used to reliably estimate subject ability from item subsets. This analysis was carried out as follows. First, for a given item subset (e.g., Easy 1), we retrieved the responses for all subjects in Experiment 1 (n=197). Next, the responses (e.g., Easy 1: 197 subjects x 36 items) were projected to the model trained in Experiment 1 (full set of TIM items: 197 subjects x 225 items). This resulted in a new set of ability scores, estimated by the model trained on the full set of items and using responses to a selected set of items (e.g., Easy 1). These steps were repeated for each item subset. Finally, the ability scores estimated from each item subset were compared to the ability scores estimated using the full TIM test (Experiment 1).

Results indicated strong positive correlations between subject ability estimated from the full test and subject ability estimated from the subsets, which ranged from (r = .58) to (r =.87) (see Figure 6).

To measure ability estimates using "Difficult" vs "Easy" subsets, the standard error of the ability estimate for each subject on each subset was plotted (Figure 6B). Standard error estimates were lower for all three difficult subsets, which suggests that these three difficult subsets provide more reliable measures of ability.

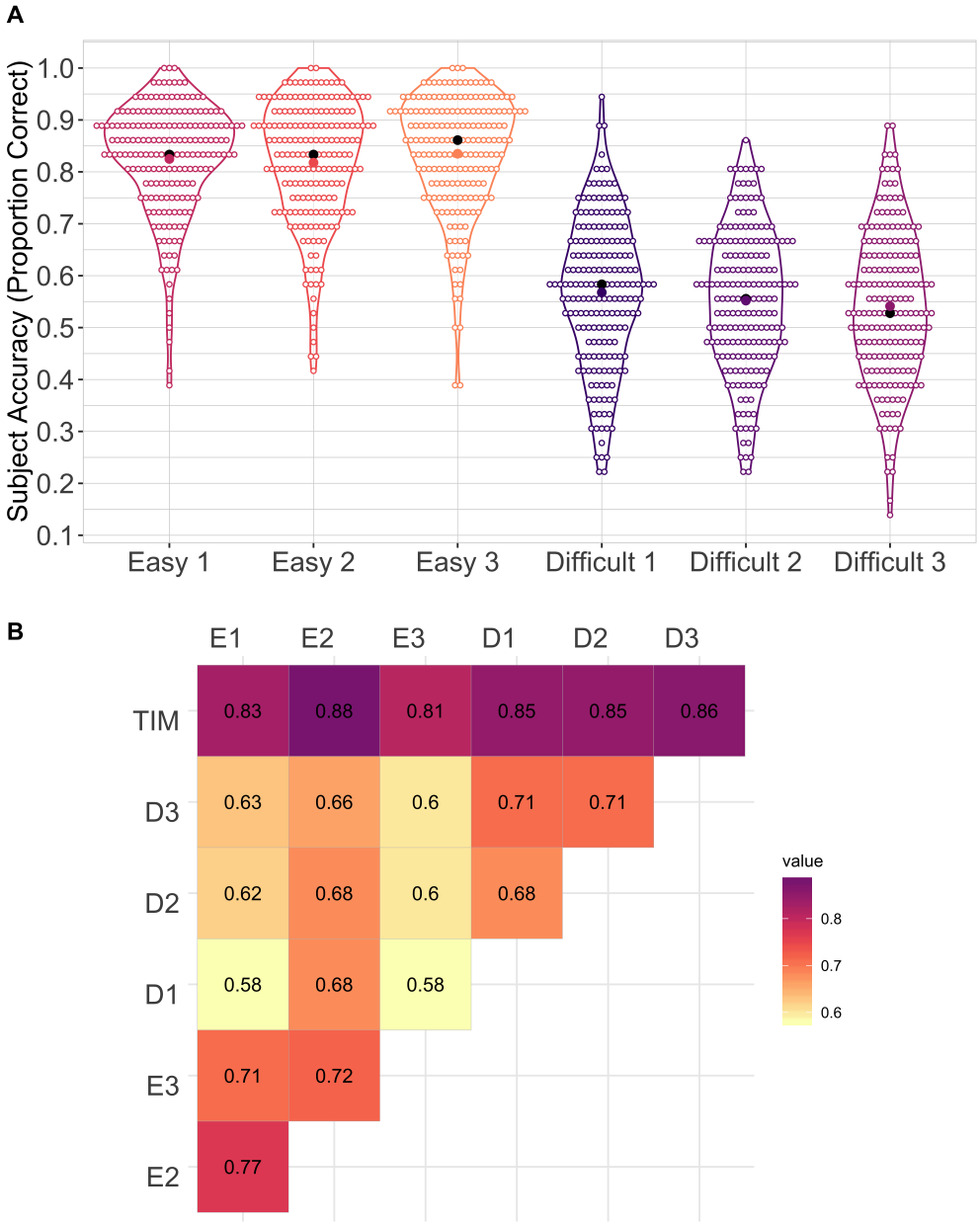

Figure 5: A) Violin plots of subject accuracy (proportion correct) on each item subset. The empty circles represent the accuracy of individual subjects on each item subset, the coloured dots represent the item subset mean, the black dots represent the item subset median. B) Pearson correlation between subject accuracy (proportion correct) on the full TIM test and all subsets. The Full TIM test was highly correlated with all six item subsets. All comparisons are significant at the 0.01 level.

4.3 Experiment 2 Discussion

Experiment 2 provides a proof of principle of the validity of sub-sampling items from the TIM test. The item subsets derived from the full TIM test provide measures of proficiency that are comparable to the measures obtained from the entire TIM test. Although proportion correct (i.e., proficiency measured as the sum of correct responses) and IRT-based ability (i.e., proficiency measured as a latent variable) yield similar comparisons in subject performance across all sets, IRT modeling provides additional information regarding the measurement error. Here, results indicated that the subsets

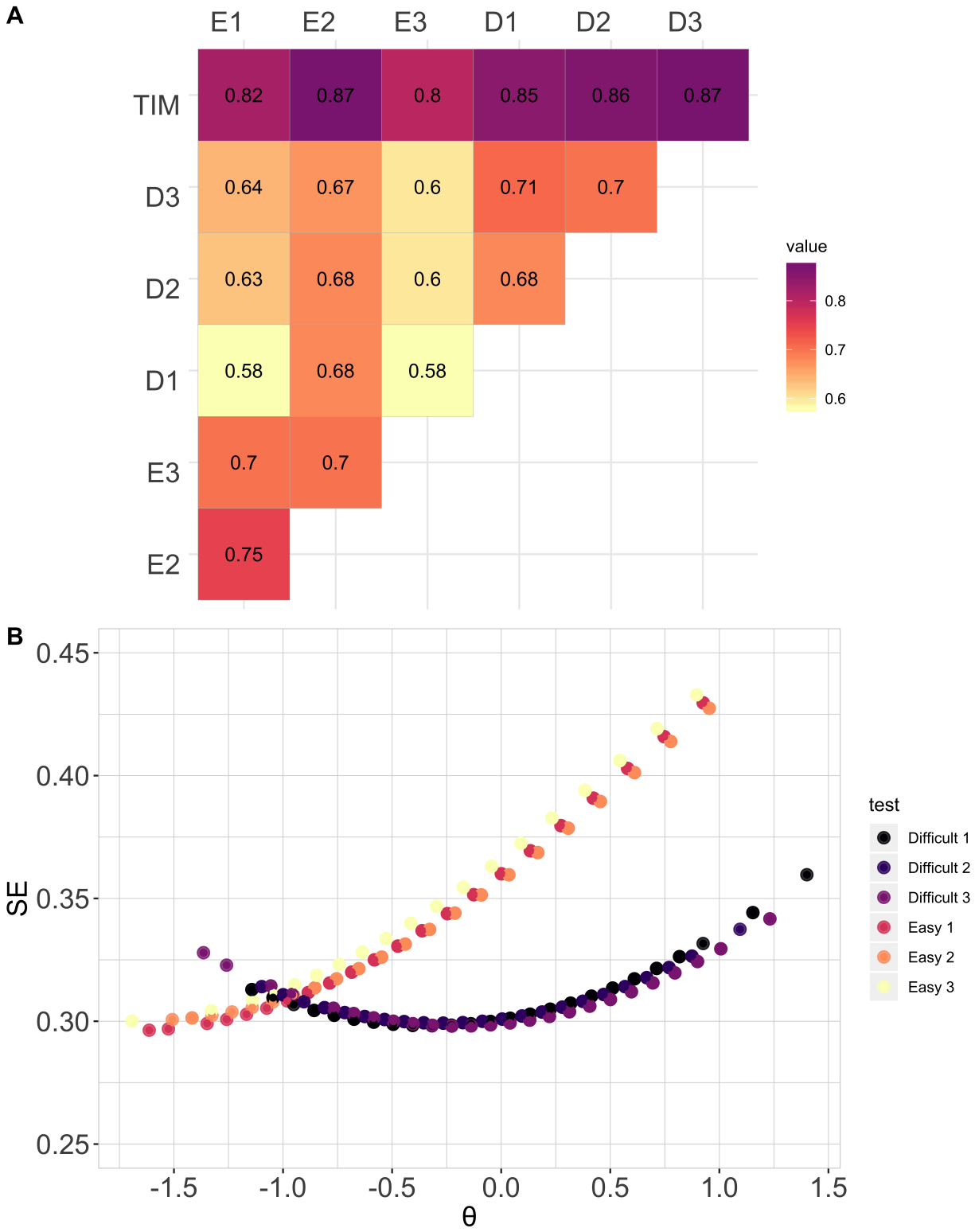

Figure 6: A) Pearson correlation between subject ability on the full TIM test and all subsets. All comparisons are significant at the 0.01 level. B) Standard error of the ability estimate for all subjects on all subsets.

provided the most reliable subject ability estimates for subjects ranging between low to average ability. In addition, the difficult item subsets yielded more reliable measures of subject ability, than the easy item subsets.

5 General Discussion and Conclusion

The objective of this study was to refine the current state of face identification testing by developing a framework for creating proficiency tests. This framework relies on IRT to calibrate item difficulty in relation to subject ability, thereby enabling the selection of subsets of items that can be combined in systematic ways to create tests of specified difficulty. These item subsets can be tailored for testing individuals of specific ability levels and for testing professionals who are busy and may only be able to spare time for short tests. Multiple tests of equal difficulty can be used also to detect *changes* in ability (e.g.,from training, experience, or age). Because items are not reused in multiple tests, proficiency improvements can be detected without confounding factors that result from repeated exposure to the same faces.

Using this framework, we introduce the TIM test, which includes items that span a range of difficulty from very easy (97% of subject endorsed a correct response to the item) to very challenging (17% of subject endorsed a correct response to the item). This range of difficulty supports the assessment of subject abilities close to random performance (accuracy of 37%) to high ability (accuracy of 89%). The TIM test was designed to address longstanding response bias issues in traditional face identification tests due to the use of rating scales and binary decision choices. Response bias poses a particularly vexing problem when comparing across groups of different ability who use the scale in different ways. The TIM test stimuli and materials that support the framework (de-identified data and code to build the student-based one parameter model) can be obtained for research use.[4] The framework and results we present provide a general foundation for future research that connects to basic theory in the psychology of face recognition, as well as to testing in research and applied scenarios.

It has become increasingly clear in the psychological literature that successful face identification requires two important skills. The first is the need to discriminate highly similar faces (i.e., "telling people apart")—long considered the basis for human expertise with faces [27]. The second, is the ability to perceive identity constantly across multiple face images that vary in appearance and image conditions (e.g., expression, viewpoint, illumination)(i.e., "telling faces together") [28, 29]. The constructed triads used in the TIM test implicitly test both skills simultaneously. In particular, the triads evaluate both the ability to "tell people together" and "tell people apart" with stimuli constructed to be challenging for both tasks.

The framework developed in this study, combined with the TIM test we introduce, provides a path for building calibrated face identification tests. To use this framework as a foundation, additional research is required to: a.) verify the stability of individual performance across time; b.) evaluate the applicability of the TIM test for other populations (e.g., non-student subjects that are not trained for forensic identification, super-recoginzers, forensic specialists); and c.) evaluate whether the TIM test predicts performance on tests more commonly used in the forensic domain. We consider each in turn.

First, a requirement of an efficient face identification proficiency test is to produce stable measures of a person's skill across time. In the absence of intervention (e.g., training), a test should produce comparable estimates of proficiency over time and from multiple subsets of items for an individual. This repeated measure over time has yet to be verified empirically for the TIM test, or to the best of our knowledge for other commonly used tests of face matching.

Second, to establish a general baseline and proof of concept, the current study was limited to a student population. Although simulation results (See supplemental) offer *prima facie* evidence that the test transferred well between students of higher and lower ability, this should be verified explicitly with other populations. Future research should focus on evaluating the TIM test for non-student populations such as forensic examiners, super-recognizers, forensic specialists, and prosopagnosics.

Third, one motivation for developing this framework concerned the challenges of measuring item difficulty in test paradigms such as identity matching. These paradigms allow for user response bias in the form of rating scales or binary choice decision options and are some of the most commonly used in forensic practice. A requirement of an efficient face identification proficiency test is to provide measures of ability that can translate to people's proficiency for applied settings. Therefore, it is important to verify that measures of proficiency gleaned from a 3AFC test, such as the TIM test, accurately predict performance in identity-matching tasks.

[4]Images used in the TIM test are available by license from the University of Notre Dame. R code to fit the Rasch model and run the analysis, de-identified subject responses, as well as PsychoPy experimental code will be made available by the UT Dallas research team at OSF website.

Finally, expecting that the TIM test can spur extensive research in the face identification community, we make the test available online to researchers. Other researchers can test individuals using the full set of items or the subsets of items used here to estimate the abilities of individuals. This supports easy comparison with the student data we report here. Using the existing model, researchers can project their data to estimate ability for other populations of interest, or they can merge their data to create a new model.

In anticipating the future of calibrated face identification tests, future research should build on the approach proposed here and examine more complex IRT models (e.g., two- and three-parameter model [30]). We based our study on the Rasch IRT model, which do not model subject guessing and assumes all items have equally discrimination parameters. More general IRT models were developed to handle both conditions. These models would offer a deeper understanding on subjects' ability and item difficult and contribute to designing well-calibrated proficiency tests. Understanding the nature of subject ability and item difficulty would offer a starting point for developing adaptive face identification tests (e.g., Computerized Adaptive Tests).

Although this study focused on face-identification, nothing in our framework is specific to facial comparisons. Researchers and practitioners can apply our work to disciplines that perform comparisons, for example, latent fingerprint, speaker, and iris identification. Our method has the potential to provide multiple forensic disciplines with the tools to create calibrate proficiency tests.

6 TIM test availability

The TIM test will be made available for research purposes, without cost, by license from the University of Notre Dame. Specifically, researchers will be able to access all TIM test images from a repository, provided by the University Notre Dame, after signing a license in which they agree to the conditions of use. This process ensures that all individuals who wish to use the test accept responsibility for adhering to subject protections and protecting subject privacy.

`https://cvrl.nd.edu/projects/data/#triad-identity-matching-tim-test-data-set`

Other materials (R code to run the analysis, de-identified data, and PsychoPy experimental code) will be made available by the UT Dallas research team at OSF website.

`https://osf.io/fd27s/?view_only=ccc063bbf9634f15859081332651e03d`

7 Appendix

IRT modelling. IRT encompasses a group of latent variable models that link item responses (e.g., face-identification judgment) to a single latent variable (e.g., face-identification ability) (cf.[31]). In the case of dichotomous items, IRT models are used to compute the probability of a correct response endorsed by the i^{th} subject on the j^{th} item. In application, an IRT model is fit to item responses, and is expressed as follows:

$$P(x_{ij} = 1|\theta_i) = c_j + (1 - c_j)g\{\alpha_j, (\theta_i - \beta_j)\}, \tag{1}$$

where x_{ij} represents the response status (1= correct, 0= incorrect) of the i^{th} subject on the j^{th}item, θ_i denotes the subject latent score (e.g.,ability), c_j denotes the item guessing parameter, α_j denotes the item discrimination parameter, and β_j denotes the item difficulty parameter. The slope of the line (α_j) determines the item's sensitivity to changes across the latent scale (the steeper, the better at discriminating subjects of different ability levels). Item difficulty (intercept) (θ_j) determines the location on the latent scale that yields a 0.5 probability of correct response. The lower asymptote of the line (c_j) is used to represent a correct answer endorsed by guessing. In this paper, we considered the Rasch one-parameter model, where α_j is constrained to a value of 1, and c_j is constrained to a value of 0. Additional IRT models are available for dichotomous and polytomous data (cf.[31]).

8 Supplemental Information

8.1 Face recognition and identification tests

Table 3: Examples face recognition tests. Task Type: “M” denotes memory task: “P” denotes a perceptual task. Tests analyzed using “IRT" appear in bold.

Task Type	Test	Stimuli	Procedure
M	**Cambridge Face Memory Test** [32]	72 items, 6 target faces	recognition (3-AFC)
M	**Cambridge Face Memory Test** (long form) [33]	102 items	recognition (3-AFC)
M	**Vanderbilt Face Matching Test** [17]	95 items, 2 catch items	recognition (3-AFC)
M	**Penn Face Memory Test** [34]	20 targets, 20 foils	old/new
P	Glasgow Face Matching Test [7]	168 image pairs	same/different (binary response)
P	Glasgow Face Matching Test (short form) [7]	40 image pairs	same/different (binary response)
P	Expertise in Face Comparison Test [13]	84 image pairs (half same identity)	same/different (response scale)
P	Person Identification Challenge Test [13]	40 image pairs (half same identity)	same/different (response scale)
P	Kent Face Matching Test [8]	200 match, 20 non-match	same/different (binary response)
P	Cambridge Face Perception Test [32]	8 upright, 8 inverted	arrange 6 images based on target similarity
P	Yearbook Test [35]	35 items	face similar to target
P	Face Identity Card Sorting Test [28]	40 images (2 identities)	group images by identities

8.2 IRT Modeling Generalizability

On average, students' face identification ability is assumed to fall at the midpoint of the face proficiency distribution. Therefore, it is unclear if IRT modeling based on a student population can generalize to other populations (e.g., high ability groups such as super-recognizers or low ability groups such as prosopagnosics). As noted, testing special populations is beyond the scope of this study. However, to simulate a scenario of generalization, we conducted IRT modeling simulations. Our goal was to examine if IRT model built from lower-ability (higher-ability) subjects provides accurate ability estimates of higher-ability (lower-ability) subjects. To do this, subject ability was estimated by projecting data from subjects of one ability group (e.g., high ability) onto an IRT model trained on a different ability group (e.g., low ability).

First, an IRT model was built using data from all 197 subjects (Model "All"). Estimated subjects' abilities served as the "ground-truth" for ability. Second, we divided subjects into a higher-ability group ($N = 99$) and lower-ability group ($N = 98$), based on their percentage of correct scores (median split). Third, we built an IRT model with the lower ability group (Model "Low") and used the model to estimate the abilities of the higher-ability subjects based on their raw responses. The projected abilities were compared to the ground truth abilities achieved from a model that included all subjects. This step simulates a real-world situation where researchers would test face-identification examiners using an IRT model that is built from student populations. Fourth, we built an IRT model with the higher-ability group (Model "High") and estimated and analyzed the abilities of the lower-ability subjects. This step simulates a real-world situation in which low-ability subjects are tested using an IRT model derived from a student population.

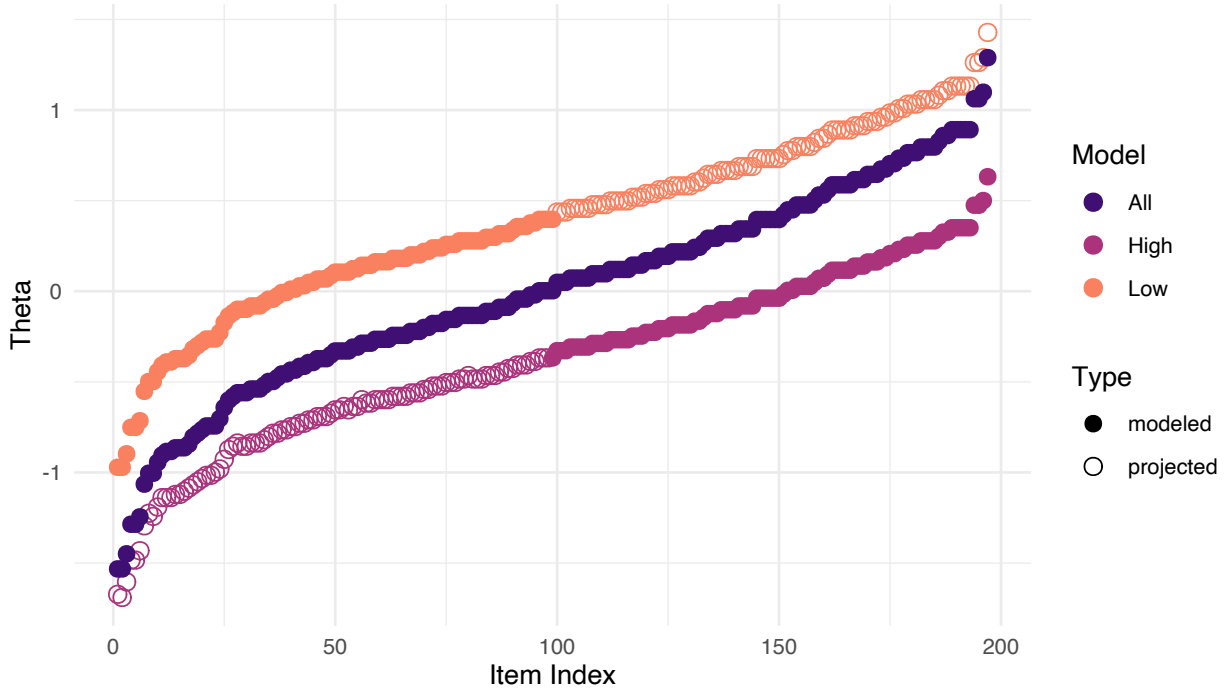

Figure 7: **IRT modeling generalization.** The middle dotted curve (purple) denotes the "ground-truth" abilities obtained from IRT models based on all subjects. The bottom dotted curve (magenta) is derived from an IRT model built from the higher-half ability subjects (determined by percent correct scores/ground-truth abilities). The top dotted curve (orange) is derived from an IRT model built from the lower-half ability subjects (determined by percent correct scores/ground-truth abilities). Filled dots denotes ability estimated when the subject is included in the model. Hollow dots denotes ability estimated through projections (i.e., when the subject is excluded in the model).

Generalizability results appear in Figure 7 and show that IRT results created from all subjects (Model "All") fall in between the model created from higher-ability individuals (Model "High") and the model created from lower-ability individuals (Model "Low"). The results can be interpreted from a correlational perspective and from the perspective of ability. From a correlational perspective, subject ability estimated from models built from all subjects correlated

positively with that achieved from models built from higher- and lower-ability subjects[5]. From the perspective of ability, estimates of ability changed, based on the samples included. When the model included samples biased towards lower-ability individuals (top curve), the estimated abilities of the higher-end subjects were higher than ground truth (middle curve). In contrast, when the model included samples biased towards the higher-ability individuals (bottom curve), the estimated abilities of the lower-end subjects were lower than ground truth (middle curve).

These results justify support the validity of generalizing IRT modeling across different subject ability levels. IRT models build from specific ability groups can accurately measure another group regardless of differences in abilities. The IRT parameters provided in the current work can serve as benchmarks for evaluating future subjects of wide-ranging ability.

[5]Correlational values across all three models is .9997, p = 2.2e-16. Notably, this value is expected as results of a direct function of IRT properties and not a novel finding based on our data.

References

[1] Eilidh Noyes and Alice J O'Toole. Face recognition assessments used in the study of super-recognisers. *arXiv preprint arXiv:1705.04739*, 2017.

[2] Meike Ramon, Anna K Bobak, and David White. Super-recognizers: From the lab to the world and back again. *British Journal of Psychology*, 110(3):461–479, 2019.

[3] Andrew W Young and Eilidh Noyes. We need to talk about super-recognizers invited commentary on: Ramon, M., Bobak, A.K., & White, D. Super-recognizers: From the lab to the world and back again. British Journal of Psychology. *British Journal of Psychology*, 110(3):492–494, 2019.

[4] Geraldine Dawson, Sara Jane Webb, and James McPartland. Understanding the nature of face processing impairment in autism: insights from behavioral and electrophysiological studies. *Developmental neuropsychology*, 27(3):403–424, 2005.

[5] Katie Marwick and Jeremy Hall. Social cognition in schizophrenia: a review of face processing. *British Medical Bulletin*, 88(1):43–58, 2008.

[6] Dana A Roark, Alice J O'Toole, Hervé Abdi, and Susan E Barrett. Learning the moves: The effect of familiarity and facial motion on person recognition across large changes in viewing format. *Perception*, 35(6):761–773, 2006.

[7] A Mike Burton, David White, and Allan McNeill. The Glasgow Face Matching Test. *Behavior Research Methods*, 42(1):286–291, 2010.

[8] Matthew C Fysh and Markus Bindemann. The Kent Face Matching Test. *British Journal of Psychology*, 109(2):219–231, 2018.

[9] NA Macmillan and CD Creelman. *Detection Theory, 2nd*. Lawrence Erlbaum Associates, New Jersey, 2005.

[10] Nicolaas Prins et al. *Psychophysics: a practical introduction*. Academic Press, 2016.

[11] Ying Hu, Kelsey Jackson, Amy Yates, David White, P Jonathon Phillips, and Alice J O'Toole. Person recognition: Qualitative differences in how forensic face examiners and untrained people rely on the face versus the body for identification. *Visual Cognition*, 25(4-6):492–506, 2017.

[12] P Jonathon Phillips, Amy N Yates, Ying Hu, Carina A Hahn, Eilidh Noyes, Kelsey Jackson, Jacqueline G Cavazos, Géraldine Jeckeln, Rajeev Ranjan, Swami Sankaranarayanan, et al. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, 115(24):6171–6176, 2018.

[13] David White, P Jonathon Phillips, Carina A Hahn, Matthew Hill, and Alice J O'Toole. Perceptual expertise in forensic facial image comparison. *Proceedings of the Royal Society B: Biological Sciences*, 282(1814):20151292, 2015.

[14] Frederic M Lord. *Applications of item response theory to practical testing problems*. Routledge, 1980.

[15] Jeremy B Wilmer, Laura Germine, Christopher F Chabris, Garga Chatterjee, Margaret Gerbasi, and Ken Nakayama. Capturing specific abilities as a window into human individuality: The example of face recognition. *Cognitive Neuropsychology*, 29(5-6):360–392, 2012.

[16] Sun-Joo Cho, Jeremy Wilmer, Grit Herzmann, Rankin Williams McGugin, Daniel Fiset, Ana E Van Gulick, Kaitlin F Ryan, and Isabel Gauthier. Item response theory analyses of the Cambridge Face Memory Test (CFMT). *Psychological assessment*, 27(2):552, 2015.

[17] A Sunday, Mackenzie, Woo-Yeol Lee, and Isabel Gauthier. Age-related differential item functioning in tests of face and car recognition ability. *Journal of vision*, 18(1):2–2, 2018.

[18] Michael L Thomas, Gregory G Brown, Ruben C Gur, Tyler M Moore, Virginie M Patt, Victoria B Risbrough, and Dewleen G Baker. A signal detection–item response theory model for evaluating neuropsychological measures. *Journal of clinical and experimental neuropsychology*, 40(8):745–760, 2018.

[19] Rafael Jaime De Ayala. *The theory and practice of item response theory*. Guilford Publications, 2013.

[20] P Jonathon Phillips, J Ross Beveridge, Bruce A Draper, Geof Givens, Alice J O'Toole, David Bolme, Joseph Dunlop, Yui Man Lui, Hassan Sahibzada, and Samuel Weimer. The Good, the Bad, and the Ugly Face Challenge Problem. *Image and Vision Computing*, 30(3):177–185, 2012.

[21] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Association*, 2015.

[22] Benjamin D Wright. Solving measurement problems with the rasch model. *Journal of Educational Measurement*, pages 97–116, 1977.

[23] I Qualtrics. Qualtrics. *Provo, UT, USA*, 2013.

[24] Jonathan W Peirce. Psychopy—psychophysics software in python. *Journal of neuroscience methods*, 162(1-2):8–13, 2007.

[25] R Philip Chalmers et al. mirt: A multidimensional item response theory package for the R environment. *Journal of statistical Software*, 48(6):1–29, 2012.

[26] Derek Beaton, Cherise R Chin Fatt, and Herve Abdi. An exposition of multivariate analysis with the singular value decomposition in R. *Computational Statistics & Data Analysis*, 72:176–189, 2014.

[27] Rhea Diamond and Susan Carey. Why faces are and are not special: an effect of expertise. *Journal of experimental psychology: general*, 115(2):107, 1986.

[28] Rob Jenkins, David White, Xandra Van Montfort, and A Mike Burton. Variability in photos of the same face. *Cognition*, 121(3):313–323, 2011.

[29] Sally Andrews, Rob Jenkins, Heather Cursiter, and A Mike Burton. Telling faces together: Learning new faces through exposure to multiple instances. *Quarterly Journal of Experimental Psychology*, 68(10):2041–2050, 2015.

[30] A Birnbaum. Some latent trait models and their use in inferring an examinee's ability. In M. Lord and M. R. Novick, editors, *Statistical theories of mental test scores*. Addison-Wesley.

[31] Dimitris Rizopoulos. ltm: An R package for latent variable modeling and item response theory analyses. *Journal of statistical software*, 17(5):1–25, 2006.

[32] Bradley Duchaine, Laura Germine, and Ken Nakayama. Family resemblance: Ten family members with prosopagnosia and within-class object agnosia. *Cognitive neuropsychology*, 24(4):419–430, 2007.

[33] Richard Russell, Brad Duchaine, and Ken Nakayama. Super-recognizers: People with extraordinary face recognition ability. *Psychonomic bulletin & review*, 16(2):252–257, 2009.

[34] Ruben C Gur, Jurg L Jaggi, J Daniel Ragland, Susan M Resnick, Derri Shtasel, Larry Muenz, and Raquel E Gur. Effects of memory processing on regional brain activation: cerebral blood flow in normal subjects. *International Journal of Neuroscience*, 72(1-2):31–44, 1993.

[35] Maggie Bruck, Patrick Cavanagh, and Stephen J Ceci. Fortysomething: Recognizing faces at one's 25th reunion. *Memory & Cognition*, 19(3):221–228, 1991.