Assessing Image Difficulty in X-Ray Screening Using Image Processing Algorithms

Anton Bolfing, Stefan Michel, Adrian Schwaninger

***Abstract*—The relevance of aviation security has increased dramatically in the last years. One of the most important tasks is the visual inspection of passenger bags using x-ray machines. In this study we investigated the role of the three image-based factors view difficulty, superposition and bag complexity on human detection of familiar prohibited items (knives) in x-ray images. In Experiment 1 we replicated earlier findings in order to provide converging evidence for the validity of these factors. In Experiment 2 we assessed the subjective perception of the same image-based factors. Twelve participants rated the x-ray images used in Experiment 1. Threat images were rated for view difficulty, superposition, clutter, transparency and general difficulty. Except for clutter ratings obtained in Experiment 2 were significantly correlated with detection performance in Experiment 1. We then developed statistical and image-processing algorithms to calculate the image-based factors automatically from x-ray images. In Experiment 3 it was revealed that most of our computer-generated estimates were well correlated with human ratings of image-based effects obtained in Experiment 2. This shows that our computer-based estimates of view difficulty, superposition, clutter and transparency are perceptually plausible. Using multiple regression analysis we could show in Experiment 4 that our computer estimates were able to predict human performance in Experiment 1 as well as the human ratings obtained in Experiment 2. Applications of such a computational model are discussed for threat image projection systems and for adaptive computer-based training.**

***Index Terms*—Airport security, image difficulty estimation, image processing, statistical model, x-ray screening.**

I. Introduction

The aim of this study is to evaluate a computational model for image difficulty assessment. Such a model could be important for applications such as Threat Image Projection (TIP) or computer based training. TIP is a software function available on modern x-ray equipment that allows inserting fictional threat items (FTIs) into x-ray images of real passenger bags. TIP is a source of motivation to screeners, provides a means of improving screeners' threat knowledge, and can be used to assess screeners' threat detection performance. Schwaninger, Hardmeier and Hofer (2004) have identified three major image-based factors influencing detection performance: View difficulty of the FTI depending on its rotation, superposition by other objects in the bag, and bag complexity. The latter comprises clutter, the texture's unsteadiness, and transparency, the relative size of dark areas. Current TIP systems project FTIs into real passenger bags based on a random ratio into a random position of the bag. As a consequence, TIP images vary substantially in image difficulty depending on effects of view difficulty, superposition and bag complexity. When TIP data is used to assess screener performance such effects make it difficult to obtain reliable measurements. The main aim of the current work is to develop a computational model using image processing in order to determine x-ray image difficulty while taking effects of view difficulty, superposition, and bag complexity into account.

The study comprises four experiments. The first experiment is a replication of earlier findings to confirm the relevance of three image-based factors in predicting human detection performance. In the second experiment we estimated the relevance of these image-based factors by correlating subjective ratings of them with the hit rate (p(hit)) in detection performance. We expect high correlations to reflect high influence of the subjectively perceived image-based factors on the measured item difficulty p(hit). In Experiment 3 we correlated the computer-based estimates with human ratings to estimate the perceptual plausibility of our computer algorithms for image-based factors estimation. Additionally, this allows us to check for possible intercorrelations among the predictors, which allows us to detect statistical dependencies among them. Finally, in experiment 4 we used multiple linear regression analyses to compare our computational model and human perception with regard to how well they can predict human detection performance.

II. Experiment I: ORT Findings Replication

A. Method and Procedure

1) Participants

The sample size of participants was twelve undergraduate students in psychology (5 females). None of the participants has had experience with x-ray images before.

Manuscript received February 24, 2006. This work was supported in part by the Zurich State Police, Airport Division.

A. Bolfing is with Department of Psychology, University of Zurich, Switzerland (e-mail: a.bolfing@psychologie.unizh.ch).

S. Michel is with the Department of Psychology, University of Zurich, Switzerland (e-mail: s.michel@psychologie.unizh.ch).

A. Schwaninger is with the Department of Psychology, University of Zurich, Switzerland (e-mail: a.schwaninger@psychologie.unizh.ch).

2) ORT Test Design

Stimuli were displayed on 17'' TFT screens at a distance of about 100 cm so that x-ray images subtended approximately 10-12 deg of visual angle. The computer program measured outcome (hit, miss, false alarm (FA), correct rejection (CR)) and the time from image onset to final decision key press.

In this study we used the X-Ray Object Recognition Test [4], which contains 256 x-ray images, half of them with an FTI (threat images), the other half without an FTI, i.e. non-threat images. Viewpoint difficulty, superposition and bag complexity are counterbalanced using the following design: 16 (threat items, i.e. 8 guns and 8 knives) x 2 (easy vs. difficult viewpoint) x 2 (easy vs. difficult superposition) x 2 (easy vs. difficult bag complexity) x 2 (threat vs. non-threat images).

The construction of the items in all image-based factor combinations as explained above was lead by visual plausibility criteria. After choosing two sets of x-ray images of harmless bags differing in bag complexity, the sixteen FTI's were projected into the bags in two different view difficulties at two locations with different superpositions each (for details see [4]).

3) Procedure

The X-Ray ORT is fully computer-based. Before starting the test, several practice trials are presented to make sure that the task is understood properly. Immediately prior to the actual test, all threat objects are presented on the screen to reduce any effects of visual knowledge about the shape of threat objects [5]. The participant's task is to decide whether a bag is OK (no threat item present) or NOT OK (threat item present). Each x-ray image disappears after 4 seconds. In addition, participants have to judge the confidence in their answer using a slider control (from "not sure at all" to "very sure"). No feedback is given to their answers. In this study, only trials containing knives have been used for analysis because of their high familiarity. This is important because we want to measure image-based factors related to visual abilities and not the visual knowledge of screeners about threat objects (for details see [5]). In 2005 the same study has been carried out with guns [7], where similar results have been obtained. A study comparing and discussing the differences is currently being conducted.

4) Statistical Analysis

A three-way analysis of variance (ANOVA) with view difficulty, superposition, and bag complexity as within-participant factors was used on percentage of detected threats (hit rate) per participant and factor combination.

B. Results

The main effects of view difficulty, superposition, and bag complexity are shown in Figure 1.

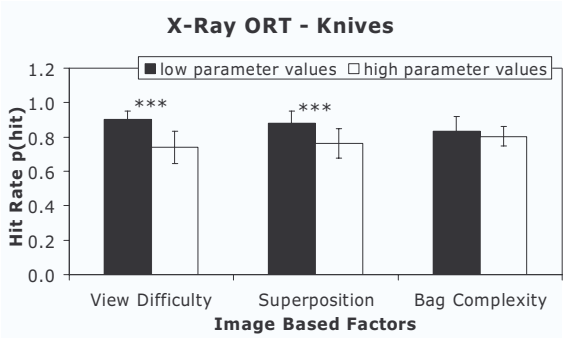

Fig. 1 Main effects of the image-based factors view difficulty, superposition, and bag complexity on hit rate in the X-Ray ORT.

The following results were obtained in the ANOVA. There were clear main effects of view difficulty, η^2=.84, $F(1,11)$=59.06, p<.001, and superposition: η^2=.65, $F(1,11)$=20.48, p<.001. The effect of bag complexity was only marginally significant, η^2=.23, $F(1,11)$=3.30, p=.10. Interaction effects: View difficulty * superposition, η^2=.19, $F(1,11)$=2.54, p=.14, view difficulty * bag complexity, η^2=.52, $F(1,11)$=11.93, p<.01, superposition * bag complexity, η^2=.01, $F(1,11)$=0.06, p= .82, and view difficulty * superposition * bag complexity, η^2=.34, $F(1,11)$=5.60, p<.05.

C. Discussion

We found clear effects of view difficulty and superposition in hit rates, which replicates the results of an earlier study using the X-Ray ORT with novices and screeners by calculating A' scores [4, 5]. Thus, the results of Experiment 1 provide further evidence for the validity of image-based factors as important determinants of x-ray image difficulty. However, the effect of bag complexity was only marginally significant whereas in earlier studies large main effects of bag complexity were found [4, 5]. This is due to the fact that in this study the dependent variable was the hit rate in contrast to our earlier studies in which A' was used. A' is a "non-parametric" measure of detection performance or sensitivity that takes the hit rate and the false alarm rate into account (for details on this and other detection measures in x-ray screening see [6]). Note that bag complexity is the only factor, which is solely dependent on the bags themselves, see section IV.B.1). Therefore, the false alarm rate is the sensitive measure for bag complexity which explains why large effects of bag complexity were found in earlier studies in which hit and false alarm rates were used to calculate A' scores [4, 5].

III. Experiment II: Image-based Factors Rating

A. Method and Procedure

1) Participants

The participants of Experiment 1 took part one week later in Experiment 2.

2) Rating

The same experimental setup was used as in Experiment 1. The participant's task in Experiment 2 was to rate the X-Ray

ORT images in terms of general difficulty and the following image-based factors: View difficulty, superposition, and bag complexity (clutter and transparency). Each of these dimensions could be rated on a 50 point scale from "very low" to "very high" using a slider control.

3) Statistical Analysis

In order to validate the influence of the subjectively perceived image-based factors on the hit rate we correlated the image-based factors ratings with p(hit) derived from Experiment 1 using Pearson's product-moment correlation.

B. Results

The correlation between objective hit rate and subjectively rated view difficulty was $r(64)=-.53$, $p<.001$, between p(hit) and superposition $r(64)=-.67$, $p<.001$, between p(hit) and clutter $r(64)=-.24$, $p=.06$, and between p(hit) and transparency $r(64)=.31$, $p<.05$.

C. Discussion

The results show clearly that there is a covariation between objective hit rates and the subjective perception of our image-based factors view difficulty, superposition, and bag complexity. This indicates a high perceptual plausibility of these image based factors. However, bag complexity, comprising clutter and transparency, shows the lowest correlation, whereas clutter does not correlate significantly with the hit rate. Again, these findings are consistent with the above mentioned observation that bag complexity, which refers to the bag content only, should be related more to false alarm rates than to hit rates.

Another explanation could be that bag complexity is difficult to rate because of low perceptual plausibility. Indeed, the two components of bag complexity (clutter and transparency) were highly correlated $r(64)=-.89$, $p<.001$. This could imply that novices have a hard time in distinguishing the two components of bag complexity and just give similar ratings for both.

IV. Experiment III: Image-based Factors Correlations between Computer Estimates and Ratings

A. Introduction

Experiment 3 was designed to develop image-processing algorithms for estimating view difficulty, superposition, and bag complexity automatically in x-ray images. These algorithms were then validated by correlating the computer-based estimates with the human ratings from Experiment 2.

B. Method and Procedure

1) Image Processing Algorithms

All image-processing algorithms developed for this purpose are based on theoretical considerations. For each image-based factor the consequences of high and low parameter values of each single image-based factor on the pixel and frequency space have been determined. Different algorithm parameters were optimized by maximizing the correlations between the image-based factor estimates and human detection performance of earlier studies [4, 5, 7].

In the following subsections the image-processing algorithms are described in turn.

a) View Difficulty

Because it is not possible to determine the degree of 3-D rotation (view difficulty) of a physical threat item solely from the 2-D x-ray image, this image-based factor is not being calculated using computational recognition algorithms, but statistically from X-Ray ORT detection performance data.

$$ViewDifficulty \quad VD_j = \left(\left(\sum_{i=1}^{n} p(hit)_i \right) - p(hit)_j \right) \Big/ (n-1)$$

Eq. 1 shows the equation for estimating view difficulty, whereas *j* denotes the index of the x-ray image in question and n is the number of bags each FTI has been projected to.

The image-based factor view difficulty is basically calculated by averaging the hit rates of a certain threat item in one of the two views presented in the ORT. In order to avoid a circular argument in the statistical model (multiple linear regression, section V.B.1) by partial inclusion of a predictor into the criterion variable, the detection performance of the one item in question is being excluded from this average detection performance estimate.

In this study, each threat item (knives only) is being displayed four times in the same view, 2 (bag complexity low vs. high) x 2 (superposition low vs. high). Therefore, the n in the view difficulty formula equals 4, but the average is calculated over the remaining three items.

b) Superposition

The image-processing algorithm for superposition simply calculates the Euclidian distance between the grayscale pixel intensities of the signal-plus-noise (SN) image and the harmless bag (N, noise) image.

$$Superposition \quad SP = \sqrt{\sum (I_{SN}(x,y) - I_N(x,y))^2}$$

Eq. 2 Image-processing formula of the image-based factor superposition, whereas $I_{SN}(x,y)$ denotes the pixel intensities of a threat image and $I_N(x,y)$ denotes the pixel intensities of the corresponding harmless bag.

For each pixel of a threat image, the pixel intensity difference between the bag with the threat item and the bag without the threat item is calculated and squared. All squared pixel intensity differences are summed up. The final image-based factor is then derived from calculating the square root of this sum of squared pixel intensity differences.

c) Clutter

This image-based factor is designed to express bag item properties like their texture unsteadiness, disarrangement,

chaos or just clutter. In terms of the depicted bags themselves, this factor is closely related to the amount of items in the bag as well as their structures in terms of complexity and fineness. The method used in this study is based on the assumption, that such texture unsteadiness can be described mathematically in terms of the amount of edges, i.e. the amount of transitions in luminosity within a certain space frequency range surpassing a certain threshold.

$$Clutter \quad CL_C = \sum_{y}^{height} \sum_{x}^{width} (I_{hp})$$

$$where \quad I_{hp} = I_N \otimes F^{-1}(HP(f_x, f_y))$$

Eq. 3 Image-processing formula of the image-based factor clutter, whereas I_N denotes the pixel intensities of the harmless bag image, F^{-1} denotes the inverse Fourier transformation and $HP(f_x,f_y)$ represents a highpass-filter in Fourier space.

We implemented this mathematical formulation by first applying on the intensity image of the empty bag a convolution kernel, which is derived from a highpass-filter in the Fourier space by inverse Fourier transformation (see Appendix). In a second step, the amount of the resulting pixels, representing edges as described above are being counted.

d) Transparency

The image-based factor transparency reflects the amount to which x-rays are able to penetrate various objects in the bag. This depends on the specific material density of these objects. Heavy metallic materials such as lead are known to be very hard to be penetrated by x-rays. For a screener, the consequence is that he cannot see any objects in front or on the back of such material.

The implementation of the image-processing algorithm for

$$Transparency \quad TR = \frac{\sum_{x,y}(I_N(x,y) < thresh)}{\sum_{x,y}(I_N(x,y) \neq 255)}$$

Eq. 4 Image-processing formula of the image-based factor transparency, whereas $I_N(x,y)$ denotes the pixel intensities of the harmless bag and *thresh* is the pixel intensity threshold beneath which the pixels are counted.

the image-based factor transparency consists in the calculation of the amount of pixels being darker than a certain threshold (<thresh) of the pixel intensity range going from 0 to 255, relative to the bags overall size (< 255, white pixels).

2) Correlations

To evaluate the perceptual plausibility of these image-processing algorithms for estimating view difficulty, superposition, and bag complexity we correlated them with the human ratings obtained in Experiment 2.

C. Results

The correlations were $r(64)=-.47$, $p<.001$ for view difficulty, $r(64)=-.44$, $p<.001$ for superposition, $r(64)=.18$, $p=.16$ for clutter and $r(64)=-.63$, $p<.001$ for transparency.

D. Discussion

Except for clutter all correlations between calculations and ratings are highly significant. Remember the high intercorrelation between the human ratings of the image-based factors clutter and transparency ($r(64)=-.89$, $p<.001$) obtained in Experiment 2. Here, in Experiment 3 we found a quite high intercorrelation between the corresponding calculated image-based factors clutter and transparency ($r(64)=-.55$, $p<.001$). Therefore, we must keep in mind that these two factors are not fully independent. This is not very surprising as we subsume them within the factor bag complexity. Nevertheless we can conclude that our calculations are compatible with the perceptual plausibility of human observers.

V. Experiment IV: Statistical Model: using Multiple Linear Regression Analysis

A. Introduction

Experiment 4 was designed to evaluate the predictive power of our computational model and to compare it to human perception as a tough baseline.

B. Method and Procedure

One multiple regression analysis was carried out using the computationally estimated image-based factors as predictors. A second multiple regression analysis was conducted using the subjectively rated image-based factors as predictors.

1) Multiple Regression Analysis

Weight estimation for a linear model can be achieved using multiple linear regression analysis whereas our image-based factors are the predictors, and the hit rate of human observers obtained in Experiment 1 is the dependent variable. Table 1 shows the abbreviations used later in this section for all dependent and independent variables.

Dependent Variable	statistical Measure	Rating
Item Difficulty	DP: Hit Rate Detection Performance	GD General Difficulty

Independent Variables	Calculations	Ratings
View Difficulty	VD_c	VD_r
Superposition	SP_c	SP_r
Clutter	CL_c	CL_r
Transparency	TR_c	TR_r

In the following, we describe the computational model and the model based on human ratings of perceived view difficulty, superposition and bag complexity.

Linear model using computationally calculated image-based factors as predictors:

$$DP = b_0 + b_1 VD_c + b_2 SP_c + b_3 CL_c + b_4 TR_c + R$$

Linear model using mean values of the subjectively rated image-based factors as predictors:

$$DP = b_0 + b_1 VD_r + b_2 SP_r + b_3 CL_r + b_4 TR_r + R$$

Criteria for comparing the different statistical models are:

1. Goodness-of-fit measures
2. Regression coefficient's significances and
3. Percentage of variance in the dependent variable (hit rate) the model is able to explain by its predictors.

C. Results

1) Computational Model

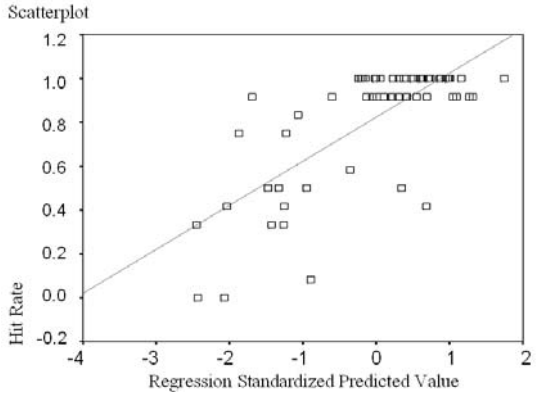

Fig. 2 Scatterplot of the multiple linear regression analysis using the calculated image-based factors as predictors with the standardized predicted values on the abscissa and the measured human performance (hit rates) on the ordinate.

In the table below the basic statistical criteria of the multiple linear regression using calculated image-based factors are listed:

Variable	B	SE B	β
VD_c	.352	.144	.288*
SP_c	.056	.013	.497***
CL_c	.000	.000	-.149
TR_c	-.172	.637	-.029

R^2=.543, $R^2(adj)$=.512, $F(4,59)$=17.49, p<0.001
* p<.05, ** p<.01, ***p<.001

2) Ratings Model

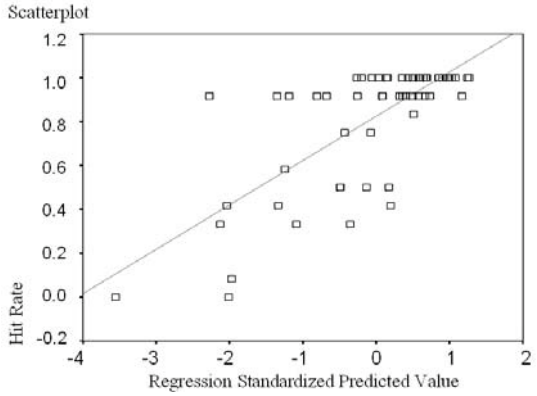

Fig. 3 Scatterplot of the multiple linear regression analysis using the subjectively rated image-based factors as predictors with the standardized predicted values on the abscissa and the measured ORT hit rates on the ordinate.

In the table below the basic statistical criteria of the multiple linear regression using subjectively rated image-based factors are listed:

Variable	B	SE B	β
VD_r	-.010	.003	-.329**
SP_r	-.021	.004	-.675**
CL_r	-.005	.007	-.131
TR_r	-.013	.008	-.338

R^2=.548, $R^2(adj)$=.518, $F(4,59)$=17.91, p<0.001
* p<.05, ** p<.01.

D. Discussion

In Experiment 4 we have developed a computational model to calculate view difficulty, superposition, and bag complexity automatically in x-ray images using image processing algorithms. These algorithms were significantly correlated with human ratings of view difficulty, superposition, and bag complexity. This result shows that our image processing algorithms are perceptually plausible. In order to benchmark our model we compared its predictive power with a model based on human perception. Our computational model using image processing algorithms for automatic estimation of view difficulty, superposition, and bag complexity could predict human detection performance in the X-Ray ORT with a correlation between model prediction and measured performance of r = .74. In order to benchmark our model we have compared it to a linear model using human ratings of view difficulty, superposition, and bag complexity. As we can see from the R^2 values of the regression analyses our computational model performed equally well as a linear model using human ratings of perceived view difficulty, superposition, and bag complexity. It is important to focus on the strength of the impacts of the single image-based factors used as predictors, the beta-weights. As we expected based on Experiment 2 and Experiment 3, the image-based factors solely depending on the empty bags contribute little to the prediction of the hit rate. In both, the computational model as well as the ratings model, only view difficulty and superposition result in significant beta-weights. This problem was discussed already in paragraph IV.D with respect to the different measures hit rate and A'. One is tempted to conclude that bag complexity with its subfactors clutter and transparency could be excluded when hit rates have to be predicted. However, it should be pointed out that in this study only knives were used and different results might be obtained for other types of prohibited items. Moreover, using another formula for estimating bag complexity it might be possible to achieve better results. In any case, there is evidence that image-based factors related to bag complexity are important to predict false alarm rates. This is even more important when TIP is activated because a high false alarm rate of a screener results in long waiting lines.

VI. General Discussion

The results show clearly, that it is possible to develop an

automatic system that calculates x-ray image difficulty using image processing algorithms. Further research needs to done regarding the division of image-based factors predicting hit and false alarm rates. Furthermore, there are great chances to enhance the existing predictors and possibly add further predictors of x-ray image difficulty. Especially in the field of bag complexity measures, probably the most challenging ones, there is some more work to be done in the future. As mentioned above, this study was conducted with knives as threat items only. Other object categories like guns, improvised explosive devices (IEDs) and other prohibited items are expected to result in different priorities of view difficulty, superposition, and bag complexity. Such studies can be highly valuable to find out how different properties of threats in terms of their shape and material can influence human detection performance and visual search. As mentioned in section II.A.3), an earlier study was carried out regarding guns as threat items [7] and further studies are planned to investigate IEDs and other threat objects in the same way. Once these data are available we will be able to discuss differences in the effects of the image-based factors in terms of different categories.

APPENDIX

Clutter formula high-pass filter:

$$HP(f_x, f_y) = 1 - \frac{1}{1 + \left(\frac{\sqrt{(f_x^2 + f_y^2)}}{f} \right)^d}$$

This formula denotes the high-pass filter as part of the clutter formula in Experiment 3, whereas f_x and f_y are its frequency components, f is its cut-off frequency and where d is its fall-off. This high-pass filter represents a 2-D matrix in the Fourier frequency-space. Therefore an inverse Fourier transform is applied to transform it into a convolution kernel in the spatial domain.

ACKNOWLEDGMENT

We are thankful to Zurich State Police, Airport Division for their help in creating the stimuli and the good collaboration for conducting the study.

REFERENCES

[1] Schwaninger, A., & Hofer, F., "Evaluation of CBT for increasing threat detection performance in X-ray screening," in *The Internet Society 2004, Advances in Learning, Commerce and Security*, K. Morgan and M. J. Spector, Eds., Wessex: WIT Press, 2004, pp. 147-156.

[2] Graf, M., Schwaninger, A., Wallraven, C., & Bülthoff, H.H., "Psychophysical results from experiments on recognition & categorization," Information Society Technologies (IST) programme, Cognitive Vision Systems – CogVis; IST-2000-29375, 2002.

[3] Schwaninger, A., "Object recognition and signal detection," in Praxisfelder der Wahrnehmungspsychologie, B. Kersten and M.T. Groner, Eds., Bern: Huber, in press.

[4] Hardmeier, D., Hofer, F., & Schwaninger, A. (2005). The object recognition test (ORT) – a reliable tool for measuring visual abilities needed in x-ray screening. *IEEE ICCST Proceedings, 39,* 189-192.

[5] Schwaninger, A., Hardmeier, D., & Hofer, F. (2005). Aviation security screeners visual abilities & visual knowledge measurement. IEEE Aerospace and Electronic Systems, 20(6), 29-35.

[6] Hofer, F. & Schwaninger, A. (2004). Reliable and valid measures of threat detection performance in X-ray screening. *IEEE ICCST Proceedings, 38,* 303-308.

[7] Schwaninger, A., Michel, S., & Bolfing A. (2005). Towards a model for estimating image difficulty in x-ray screening. *IEEE ICCST Proceedings, 39,* 185-188.