JOURNAL OF INFORMATION SYSTEMS APPLIED RESEARCH

Volume 16, No. 3
November 2023
ISSN: 1946-1836

In this issue:

4. **Optimizing a Convolutional Neural Network Model in Amazon SageMaker for an Autism Detection Tool, EZ Autism Screener**
Catherine M. Ata, City University of Seattle
Sam Chung, City University of Seattle
Brian Maeng, City University of Seattle

23. **Are Companies Responsible for Internet of Things (IoT) Data Privacy? A Survey of IoT User Perceptions**
Karen Paullet, Robert Morris University
Adnan A. Chawdhry, Pennsylvania Western University
Jamie Pinchot, Robert Morris University

33. **E-Commerce Drone Delivery Acceptance: A Study of Gen Z's Switching Intention**
Jeffrey P. Kaleta, Appalachian State University
Wei Xie, Appalachian State University
Charlie Chen, Appalachian State University

45. **The Effect of Mental Illness on Compensation for IT Developers**
Alan Peslak, Penn State University
Wendy Ceccucci, Quinnipiac University
Kiku Jones, Quinnipiac University
Lori N. K. Leonard, University of Tulsa

58. **Measuring Learners' Cognitive Load when Engaged with an Algorithm Visualization Tool**
Razieh Fathi, Smith College
James D. Teresco, Siena College
Kenneth Regan, University of Buffalo

68. **Virtual Reality in Special Education: An Application Review**
Yi (Joy) Li, Kennesaw State University
Zhigang Li, Kennesaw State University

iscap
Information Systems & Computing Academic Professionals

The **Journal of Information Systems Applied Research** (JISAR) is a double-blind peer reviewed academic journal published by ISCAP, Information Systems and Computing Academic Professionals. Publishing frequency is three to four issues a year. The first date of publication was December 1, 2008.

JISAR is published online (https://jisar.org) in connection with CONISAR, the Conference on Information Systems Applied Research, which is also double-blind peer reviewed. Our sister publication, the Proceedings of CONISAR, features all papers, panels, workshops, and presentations from the conference. (https://conisar.org)

The journal acceptance review process involves a minimum of three double-blind peer reviews, where both the reviewer is not aware of the identities of the authors and the authors are not aware of the identities of the reviewers. The initial reviews happen before the conference. At that point papers are divided into award papers (top 15%), other journal papers (top 30%), unsettled papers, and non-journal papers. The unsettled papers are subjected to a second round of blind peer review to establish whether they will be accepted to the journal or not. Those papers that are deemed of sufficient quality are accepted for publication in the JISAR journal. Currently the acceptance rate for the journal is approximately 35%.

Questions should be addressed to the editor at editor@jisar.org or the publisher at publisher@jisar.org. Special thanks to members of ISCAP who perform the editorial and review processes for JISAR.

JOURNAL OF INFORMATION SYSTEMS APPLIED RESEARCH

Optimizing a Convolutional Neural Network Model in Amazon SageMaker for an Autism Detection Tool, EZ Autism Screener

Catherine M. Ata
atacatherine@cityuniversity.edu

Sam Chung
chungsam@cityu.edu

Brian Maeng
maengjooyol@cityu.edu

School of Technology & Computing
City University of Seattle

Abstract

Autism is a neurological and developmental disability caused by changes in the brain's development that also affects the facial tissues. Thus, children with autism show distinct facial features that are not present in average children. Studies reveal increasing prevalence of autism; however, acquiring affordable, trouble-free, and practical early screening tools is a current concern. This impacted early detection and diagnosis of autism which also influenced effective intervention. How can we employ innovative technology, like computer vision and deep learning to build an inexpensive and universally accessible autism screener to prevent late detection and diagnosis of autism? How can we enhance public access to this screening tool and minimize the difficulties involved in the assessment process? We built a basic Convolutional Neural Network (CNN) binary image classifier with seven (7) layers including the input and output layers. This initial model produced positive outcomes with a specificity score of 90.38% and this is the most important evaluation metric for health-related problems like screening for autism. We optimized this model by performing hyperparameter tuning using a cloud machine learning platform, Amazon SageMaker. The tuning job also produced a superior and robust model as reflected in the F1 score of 94.74%. It correctly classified 95% of the images. The model's specificity indicates it correctly identified 100% of those without autism as non-autistic; the recall indicates it correctly identified 90% of those with autism as autistic while its precision indicates a 100% probability that those identified by the model as autistic have autism. Tuning this model took 6 minutes. We integrated this model into a simple iOS application for mobile devices.

Keywords: autism screener, autism detection, autism screening, autism facial recognition, AWS SageMaker image classifier, CNN image classifier

1. INTRODUCTION

The main objective of this research project is to improve access to initial screening for autism or autism spectrum disorder (ASD) and to minimize the difficulty of the overall assessment process. To accomplish this, a Convolutional Neural Network (CNN) binary image classification model is built from scratch as the backend for a simple and user-friendly iOS application for mobile

devices such as iPhones and iPads. This will later be expanded to Android devices to cast an even wider net.

Problem Statement

With the increasing prevalence of autism worldwide and limited access to screening tools, especially in developing countries, countless cases are detected and diagnosed late; it is also highly likely that many children with autism are undiagnosed and untreated. The negative symptoms related to this disorder may worsen and can lead to life-long problems related to developmental, learning, communication, and social abilities, and may even cause premature death (Yin, Mostafa, & Wu, 2021). Early detection is crucial, so necessary support and treatment will be provided. Most screening and diagnostic tools for autism currently available are expensive and are normally done in the clinical setting, thus requiring regular doctors' or specialists' appointments. It is not only costly but also a tedious process. This has been an ongoing issue that impacts both individuals with autism and their families. Extensive research has been conducted to resolve the assessment and intervention issues; however, it is apparent that with the rising number of cases, there is still a large need for further research (Koegel, Koegel, Ashbaugh, & Bradshaw, 2014).

This issue needs to be addressed, and based on previous and ongoing studies, we can take advantage of advanced applied science and machine learning or deep learning.

In our research, we determined how to employ a deep learning model as an autism screening tool which can be accessed by the population to bridge the gap between the resources that are presently available to the public and have a more practical, accessible, and cheaper screening tool. Our goal is to improve early detection and diagnoses and reduce undetected cases of autism in children.

Motivation

Despite ongoing extensive research, increasing assessment and intervention facilities, as well as access to support groups, there is still an existing gap in availability of these resources (Durkin, Elsabbagh, Barbaro, Gladstone, Happe, Hoekstra, Lee, Rattazzi, Stapel-Wax, Stone, Tager-Flusberg, Thurm, Tomlinson, & Shih, 2015). In fact, those resources are not easily accessible to everyone who needs them. This is mainly because of two reasons. First is the overall access; in the United States and other developed countries, they are mostly available in urban and suburban areas and are hard to access for someone from remote areas. This inconvenience may discourage those groups from initiating and seeking proper help, starting with screening for autism; in underdeveloped countries, these resources are scarce. Second is the cost and effort involved; both informal screening and formal diagnostic testing cost money. Although informal screening costs less, it still requires time and effort, while formal diagnostic testing is approximately ten (10) times more expensive, plus there is the time and effort involved.

Inaccessibility and the cost further hinder timely identification and intervention of autism in the population. As mentioned, this is of paramount importance. It can dictate the success of interventions for improving the quality of life of those with autism as well as helping families cope with the difficulties involved (Badzis & Zaini, 2014). They highlighted the importance of early detection of young children with ASD and finding remedies to help erase or minimize the symptoms and complications of this condition. The primary advantage of early identification of young children with ASD is for parents, teachers, and other people to produce strategies on how to deal with autistic kids. Koegel et al. (2014) also emphasized the importance of early detection and intervention of ASD in young children. They also suggested that the sooner the intervention is started, the better the outcome will be. Their study also discussed the short- and long-term benefits of early diagnosis and intervention. The major benefit is that ASD does not have to be a life-long disability. With early interventions, as reported by research clinics, most children can attend regular education classrooms, and some of them have lost the diagnosis. Early intervention reduces the need for more major and expensive interventions later, thus minimizing cost.

There is well-documented knowledge that children or individuals with autism share the same distinct facial features, as mentioned in the papers by Ahmed, Aldhyani, & Jadhav, 2022; Aldridge, George, Cole, Austin, Takahashi, Duan, & Miles, 2011; Beary, Hadsell, Messersmith, Hosseini, & Soltanian-Zadeh, 2022; Lu & Perkowski, 2021; Rahman & Subashini, 2022; and Sewani & Kashef, 2020. There are also several screening methods used to detect autism. However, we can leverage the advancement in technology such as image processing tools and computer vision. These can be used to process, analyze, and understand facial images of individuals with or without autism. We can also implement deep neural networks, particularly CNNs, to infer and identify if the individual in the image has autism or not (Beary et al., 2022).

Approach
Most published research studies employed transfer learning and hybrid approaches. Between the two approaches, transfer learning, specifically with MobileNet, proved to be the most promising. However, it is highly possible to build our own CNN model for the task and improve its performance by taking advantage of cloud web services for machine learning.

This work is harnessing a unique approach to optimizing the performance of a freshly built, with a considerably basic architecture, CNN model, through Amazon SageMaker hyperparameter tuning. The model is initially built, trained, and evaluated using Python and TensorFlow. The Kaggle ASD Facial Images dataset used for training the initial model is uploaded to Amazon S3 that is reformatted to a file format accepted in Amazon SageMaker. Using a script mode approach, the same model is trained in Amazon SageMaker with the reformatted image dataset. This is optimized by creating a tuning job in Amazon SageMaker. This is again fed with the reformatted dataset and is provided with ranges of chosen hyperparameters. The best-tuned model is then deployed to a mobile application for iOS devices.

Conclusions
The initial model without retraining and hyperparameter tuning provided good evaluation results. However, after automatic tuning in Amazon SageMaker, the best-tuned model outperforms the initial model, and has the same accuracy as the MobileNet used in one of the related works. With Amazon SageMaker, the tuning job takes less amount of compute time. The tuned model can also be stored in the registry, and then this can be used for similar tasks later. This model can also be retrained and tuned further using the AWS environment with less compute time.

2. BACKGROUND

Autism or autism spectrum disorder (ASD) is a neurological and developmental disability that affects how individuals interact with others, communicate, learn, and behave due to differences in the brain, which is common in most genetic conditions. Symptoms commonly appear during the first two years of birth, but they can be diagnosed at any age (NIMH, 2022). Based on numerous studies, it is believed that ASD has multiple causes and continuous studies are done to learn more about them and their impact on those with ASD. Despite all these studies, it continues to be a challenge. There has been an increasing prevalence of ASD over the years. Based on the current data from the Centers for Disease Control and Prevention (CDC), about 1 in 44 children in the United States has ASD. It occurs in all racial, ethnic, and socioeconomic backgrounds, and is known to be more common among boys than girls (CDC, 2022). And based on the World Health Organization (WHO) records as of March 2022, 1 in every 100 children worldwide has autism.

Studies suggest that permutations of genetics, environment, or the interlinkage of both changes the embryonic developmental patterns that causes alterations in the brain, which is intimately tied to the development of facial tissues. This alteration in the embryological brain results in autism. The brain is the foundation on which the various parts of the developing face grow; thus, changes to the developing brain, as we see in autism, suggest that the development of the faces of children with autism may reflect subtle facial differences compared to typically developing children (Aldridge et al., 2011).

Children or individuals with autism or ASD do not tolerate physical contact, making it difficult to perform noninvasive but direct quantitative measurements of the body. However, those with autism present similar unusual craniofacial characteristics, or dysmorphic skull and facial features, such as an unusually broad upper face with wide-set eyes, shorter middle region of the face including cheeks and nose, and broader or wider mouth and philtrum – the divot below the nose, above the top lip (Aldridge et al., 2011; Beary et al., 2022; Rahman et al., 2022). These distinct signs of autism can be useful for early recognition and detection of the disorder with innovative technology like artificial intelligence (AI), machine learning (ML), and deep learning (DL). Deep neural networks (DNN), particularly convolutional neural network (CNN) models, are known to be highly exceptional when it comes to providing solutions that require image-based and video-based analysis as well as pattern detection or recognition. That is why several works adopted transfer learning with pre-trained CNN models like EfficientNet, Inception V3, MobileNet, VGG-16, and Xception that tackled the same problem.

3. RELATED WORK

These explorations and experimentations covered in the literature review embarked upon using different forms of AI to help detect and screen for autism at an early stage. This section covers a discussion on why and how we can apply AI and

DL models as screening tools, and the different approaches to solving the problem at hand.

Literature Review

Multiple research projects tackle a similar problem, supporting the benefits of leveraging image-based classification and/or facial recognition models, particularly convolutional neural networks (CNNs). Ahmed et al. (2022) concluded that CNN image classification models are useful for early detection and diagnosis of autism by extracting those distinctive facial features from facial images and then classifying them. According to Beary et al. (2022), facial analysis can best provide early detection and diagnosis of autism based on distinct facial features. Lu et al. (2021) concluded that DL models as viable, easy, and accurate screening solutions for autism in children. Research by Rahman et al. (2022) also supports why we can use ML or DL as a screening tool for autism; according to them, CNN models are excellent at detecting hidden patterns and extracting features from colored 2D or 3D images. We can also take advantage of hybrid approaches like in the studies by Sewani et al. (2020) and by Yin et al. (2021). Sewani's work combined standard ML and DL for analyzing and classifying images of the brain taken from functional magnetic resonance imaging (fMRI) for better output. Yin's research used a similar approach by using traditional ML and AI methods like an autoencoder (AE) for advanced feature extraction on brain images produced by fMRIs.

Looking further at these research projects, they employed different approaches like transfer learning and hybrid approach as mentioned. The work by Ahmed et al. (2022) utilizes transfer learning of pre-trained CNN models like MobileNet, Xception, and Inception V3, then retrained each model on the same Kaggle ASD Facial Image dataset to extract and analyze distinctive autism facial features. Among the pre-trained models, MobileNet outperformed the rest with 100% training and 95% validation accuracy scores at 35 epochs. Interestingly, they deployed this model in a web-based app for autism detection. Similarly, Beary et al. (2022) used MobileNet and added fully connected dense layers for facial analysis and image classification to categorize images of autism vs. non-autism from the Kaggle ASD Facial Image dataset. This produced a highly performing transfer learning model with a 94.64% accuracy score at around fifteen (15) epochs on the test data. In Lu et al. (2021) research, they made use of multiple datasets: the Kaggle ASD Facial Image and East Asia ASD Children Facial Image, and retrained VGG-16 on these datasets separately and then combined the datasets. Based on that, they found that racial factors play a significant impact on the performance of the model. The results using Kaggle dataset showed 51.3% accuracy and 66.7% F1-score; using the East Asian dataset showed 95% accuracy and 95% F1-score; lastly, using the combined dataset showed 23.9% (East Asian) FP rates. Again, in the research by Rahman et al. (2022), training was done on the Kaggle ASD dataset and extracted distinct features on facial 2D images of children. They performed several experiments, training multiple pre-trained CNN models such as MobileNet, Xception, and EfficientNet as the feature extractors, and attached a deep neural network (DNN) binary image classifier to each. This yielded results with the Xception as the best performing model with 88.46% sensitivity, 91.66% specificity, 88% NPV, 92% PPV, and 96.63% AUC. On the other hand, there is a study that resorted to automatic encoding combined with computer vision for face detection and emotion and attention analysis in children (Egger, Dawson, Hashemi, Carpenter, Espinosa, Campbell, Brotkin, Schaich-Borg, Qui, Tepper, Baker, Bloomfield, & Sapiro, 2018). The paper by Egger's group only discussed their approach at a high level and used live streamed or recorded videos collected in a clinical setting in their study. The hybrid approach done in Sewani's group experimented on standard ML models like K-nearest neighbors (KNN), Support vector machine (SVM), and Random Forest (RF), and each was attached with AEs. They also experimented on CNN they built from scratch with AEs and k-fold cross-validation. All of these were trained using ABIDE (Autism Brain Imaging Data Exchange) dataset with fMRI. Another hybrid method that also made use of the ABIDE dataset was done in the research by Yin's group. They applied an AE to extract advanced features and then trained a newly built deep neural network (DNN) on it. Then they incorporated the pre-trained AE with the DNN model and trained it with raw features found in the fMRI images. The latter produced better accuracy and ROC AUC scores, 79.2% and 82.4%, respectively.

They all reinforced the documented claims that deep learning models can provide a viable, easier, and cheaper screening solution for autism in children.

There is another research that addressed a closely similar problem but focused more on genetic syndrome in general, wherein ASD is one of them (Hong, Zheng, Xin, Sun, Yang, Lin, Liu, Li, Zhang, Zhuang, Qian & Wang, 2021). The research conveyed that many genetic syndromes

have unmistakable facial dysmorphias. VGG-16, a facial recognition model, was implemented to screen for genetic syndrome in children. The model's performance and outputs were compared to the professionals' screening process, and one of these comparisons was made between the VGG-16 model and a senior pediatrician with genetics training experience, yet VGG-16 outperformed the pediatrician.

Review Conclusions

The related works included here support the theory mentioned in this research that early detection of ASD will help in determining the causes and remedies for young children with ASD, and this can also increase the success of the interventions. They underpin the hypothesis that implementing innovative technology on image-based and/or video-based facial analysis like computer vision, forms of advanced automatic algorithms, and convolutional neural networks (newly built or pre-trained) can be adopted to aid early detection and diagnosis of autism. They attest that improving the accessibility of screening and diagnostic tools is highly valued, and we are proposing that with the use of deep learning and computer vision, these tools can be available to the public, which they can access through their own mobile devices. All these previous studies boost and warrant the efforts of this research.

Thus far, transfer learning on MobileNet by Ahmed, Aldyahani, and Jadhav has been the most solid CNN model for this problem, using a similar facial images dataset, with 95% accuracy. The CNN model with autoencoder using the fMRI images dataset by Sewani and Kashef also looks promising with 84.05% accuracy, 80% sensitivity, and 75.3% specificity.

Two of the studies mentioned built a new CNN model from scratch and attached it with an under-complete autoencoder to extract key features to feed to the CNN model.

Based on the results of transfer learning and hybrid approaches, we can say that building a simple CNN model can be as effective. This can also be improved with automatic model tuning or hyperparameter tuning using Amazon SageMaker.

So far, there are no documented similar research studies that have leveraged on cloud-based ML platform. This separates our research from the ones previously mentioned. There is also not a publicly and easily available autism screening tool that everyone can use, which is also the objective of our research. In addition, the best-tuned model is as accurate as the MobileNet. The best model was evaluated against other metrics, and it outperformed all the other models mentioned in other research.

Appendix A has a synthesis matrix showing a comparison of the approaches used in the different related works and in this research study.

4. APPROACH

Requirements

Several items should be in place prior to starting development. Among them are the following:

- Google Colab Pro account for unlimited access to available GPU accelerator.
- AWS account to use Amazon S3 and Amazon SageMaker web services.
- Apple account to use XCode IDE and create an iOS developer account.
- iOS developer account to be able to run the app on actual mobile devices.
- Facial images dataset with appropriate labeling: Autistic and Non-autistic.
- CNN model as the image binary classifier.

Design and Implementation: Initial Model

The first stage of development focuses primarily on the building, training, and evaluation of the model once the appropriate and good-quality image dataset is available. The initial CNN model was built in the Google Colab environment using Python programming, necessary libraries like NumPy, Pandas, Matplotlib, and ML frameworks like TensorFlow and Keras.

During the building and training stage, some of the parameters that were explored are the different layers such as the Conv2D, MaxPool, DropOut, BatchNorm, Flatten, and Dense layers. In this model, we primarily used the Conv2D to extract the features by filtering the images, and with the help of the rectified linear unit (ReLU) activation function, patterns and specific features were detected; the MaxPool layers that follow each Conv2D layer helped in condensing the image to enhance the features extracted; the DropOut layer is used to prevent overfitting by arbitrarily removing some of the units in the neural network.

Figure 1 shows the architecture of the simple CNN model for both the initial and tuned models. The model has a total of seven (7) layers including the input and the output layers. As we can see, the first Conv2D is the input layer, in this architecture, we used 32 filters with the input shape set to 100 x 100 x 3 (height, width, and

color channels – RGB), a 2 x 2 kernel size, a ReLU activation function, and the padding is set to 'same' to ensure that the convolution operation is also performed in the border values. Then we added a MaxPool layer to reduce the spatial dimension of the images. The first hidden layer is composed of a Conv2D and a MaxPool layers, while the succeeding two (2) hidden layers are composed of a combination of Conv2D, MaxPool, and Dropout layers. We increased the number of filters for each Conv2D, from 64, 128, and 256. They have the same kernel size, activation function, and padding parameter settings. The DropOut layers' setting for the last two (2) hidden layers is 20%. The last DropOut layer is followed by a Flatten layer to convert the 2D array of the image feature map into a 1D flattened matrix, which is fed to the fully connected Dense layer with 512 neurons and ReLU activation function. The flattened matrix is then passed to the output layer which is a Dense layer with Sigmoid activation function (best suited for binary classification) and with 2 units or neurons wherein each corresponds to a class name or label.

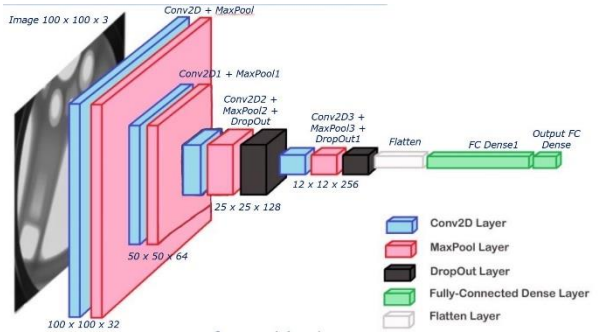


Figure 1: CNN model architecture

When training the model, to further avoid overfitting, we applied early stopping with a patience value of two (2), to ensure that after 2 epochs, if the model performance does not improve, the training is terminated. Other hyperparameters applied during training include batch size, learning rate, and the number of epochs. Although the number of epochs does not seriously affect the model's performance with early stopping, it ensures we provide enough training for the model. In this experiment, the best batch size and learning rate combination are 50 and 0.0001, respectively. We also experimented with the image size, which also impacts the model's performance. For this model, the best image size is 100 x 100.

Appendix B contains the summary of information about the CNN model, which includes the layers, the output shape of each layer, the number of weights (parameters) in each layer, the total number of parameters of the model, as well as the trainable parameters.

After training, we saved the model's weights and architecture, and then loaded the same model to perform the evaluation using a separate test data. We used several evaluation metrics to check the performance of the model. The Findings section further discussed the different evaluation metrics as well as the model's results.

Amazon SageMaker Model

Before deploying the CNN image classifier, it is necessary to ascertain that we use the best model. Thus, the second stage is geared to retraining and hyperparameter tuning which is done in a different environment; in this case, we used Amazon SageMaker, specifically, its automatic tuning or hyperparameter tuning feature.

Amazon SageMaker, as one of the cloud services offered in the AWS ecosystem, offers potential solutions to vast Data Science and machine learning projects. Some of the features it offers include access to data and pre-trained models, data analysis, building and training custom models, hyperparameter tuning, and deploying high-quality ML models at the least compute time, thus improving productivity ten (10) times more (AWS, 2022). It also offers labeling jobs for all kinds of data, including images.

Model training and hyperparameter tuning in Amazon SageMaker is known to take less time and cost, and it is highly scalable without having to manage infrastructure. The best part about hyperparameter tuning using SageMaker is it automatically adjusts thousands of algorithm parameter combinations to find the most accurate predictions. This can reduce the time and effort to get the best-performing model compared to doing it manually. It also provides built-in tools, including ML frameworks like TensorFlow, Pytorch, MxNet, and Scikit Learn, as well as open-source libraries like TensorBoard and so much more, which allows full model customization (AWS, 2022).

This is another involved process. To train the same model in SageMaker, we used a script mode approach. This is done by creating an endpoint which is the Python script that contains definitions of the model's architecture and with similar parameters used. Each parameter is initially set to lower default values. For instance, we decreased the number of epochs to ten (10), we set the batch size to thirty-two (32) and increased

the learning rate to 0.001. We converted the image dataset into a compressed .npz format before using it for retraining and hyperparameter tuning of the existing model. Then, we stored the model and the compressed image dataset in Amazon S3. To further optimize the model, we performed automatic hyperparameter tuning. We added a range of values for each hyperparameter which includes the number of epochs, the learning rate, the batch size, and the optimizer. Finally, we converted the best model into a CoreML format (.mlmodel) to deploy it in an iOS mobile application.

Appendix C shows the summary of the Amazon SageMaker CNN model which is also equivalent to the initial CNN model. The total and trainable parameters are the same as the numbers found in the initial CNN model.

iOS Mobile Application

The third stage is developing the mobile application. It is a prototype with a single page that allows taking a facial image and using the integrated best-performing CNN model to identify if there is autism or not.

Appendix D shows building of the application with the deployed best autism model in XCode IDE and running it on an actual iPhone device.

Figure 2 shows a sample image taken when using the app in an actual iPhone device.

Figure 2: Taking a sample image in an actual iPhone

Technologies Used

The technologies that are used when building the entire research project include the following:

- Google Colab with GPU accelerator using Python programming and ML frameworks like TensorFlow and Keras for building, training, and evaluating the initial CNN model
- Amazon S3 for data and model storage
- Amazon SageMaker ML web service for model hyperparameter tuning
- CoreML Apple framework to access the CNN model from Amazon SageMaker and then integrate it into the iOS app.
- XCode IDE combined with Swift programming for building the iOS app

Figure 3 represents the set of the main tools used in this approach.

Figure 3: Technologies used

5. DATA COLLECTION

The dataset used for this research was originally accessible in Kaggle but can be retrieved from GitHub with the link provided under the References section. The Kaggle ASD Facial Image dataset has three (3) subsets, the training subset with 1,327 images belonging to the Autistic class and another 1,327 images belonging to the non-Autistic class; the validation subset with only forty (40) images for each class; and the test subset with 140 images for each class. However, to increase the size of the dataset, we added more images to the training subset, which we retrieved from the results of a Google search. This gave us a total of 1350 images for each class of the training subset.

Most of the images are in colored scale while the rest are in grayscale. The images also have varying sizes, facial orientation, quality, and fidelity. The images only show the faces of the children who are both boys and girls but there are more images of boys than girls with a 3:1 ratio for the autistic class and a 1:1 ratio for the non-autistic class; the children in the images belong to the age range between 2 and 14 years old, with most of them from age 2 to 8 years old. As for

the race distribution, there are more Caucasian children than those of color, about a 10:1 ratio.

In this experiment, we used the updated training set, and the test set of the Kaggle ASD dataset for training and evaluating the initial and Amazon SageMaker model. We further divided the training set into training and validation sets into 80:20 split ratio, 80% for training, and 20% for validation. We did not utilize the images in the original validation subset as there are images that are duplicates of some of the images in the train subset, and this may affect the model's performance.

6. DATA ANALYSIS, VISUALIZATION, AND PREPROCESSING

Data Analysis

It is important to ensure that we have an excellent quality dataset after data collection. We performed basic analysis, like checking the count for each class to make sure there is equal distribution of the classes. We also must ensure each image has the correct shape, i.e., 100 x 100 x 3, where each value corresponds to the height, width, and color channels. We also checked that the number of labels matched the number of images for each class.

Data Visualization

Figure 4 shows random samples of images from the training dataset taken from GitHub with their original text labels while Figure 5 shows random samples of images after converting the original labels to numeric labels while training in Amazon SageMaker.

Figure 4: Sample images with original labels

Figure 5: Sample images with numeric labels used in Amazon SageMaker

Data Preprocessing

All images are set to a colored scale with three color channels: red, green, and blue (RGB). Rescaling the images is also necessary as this normalizes the RGB values of each pixel in the images to minimize the computing time required. Another important part of image preprocessing is ensuring all images have uniformity in size and are then resized to 100 x 100 pixels while considering the best resolution for this model as well as the compute power required.

Most machine learning and deep learning models are trained with a large amount of data. Since there is only one publicly available dataset for this specific research problem, the dataset size may not be sufficient. Data augmentation is convenient to use when it comes to increasing the data size. Since we are dealing with facial images, extra consideration is taken when it comes to the type of data augmentation that can be used effectively. We experimented on different data augmentation techniques like random crop, random rotation, random contrast, random flip, and random zoom. However, for this specific model, we only utilized a horizontal random flip. When using other techniques, they significantly impacted the performance of the model, making it less accurate.

7. FINDINGS

Initial Model

After multiple experiments on the different hyperparameters and among all the models built, trained, and evaluated, the performance of the CNN model that we used in this project is quite promising. Since the dataset is equally distributed, we used the accuracy score, which is the proportion of correct predictions, and the loss score, which is the prediction errors, as evaluation metrics. There are few models with better accuracy and loss scores, however, we observed overfitting and underfitting. Figures 6 and 7 show the model's training and validation accuracy and loss, respectively. Although the

numbers are not magnificent, we can see that this model will generalize better with very minimal overfitting.

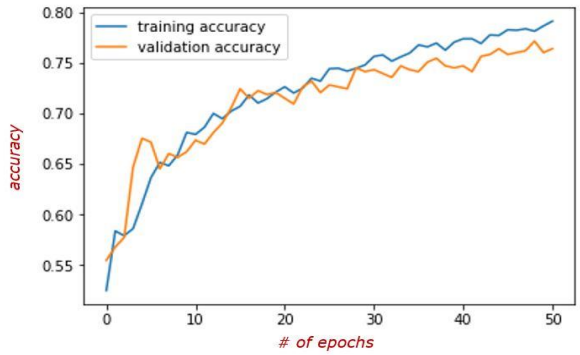


Figure 6: Training and validation accuracy scores

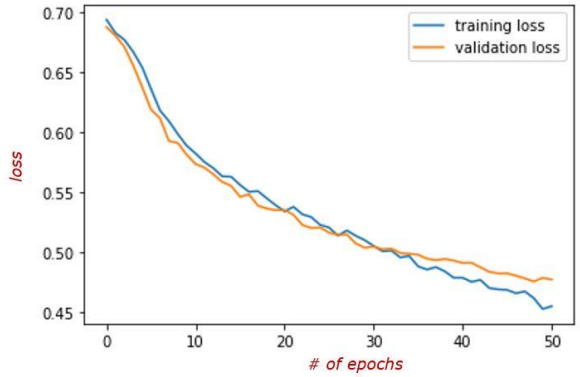


Figure 7: Training and validation loss scores

Appendix E shows the results of the prediction produced by the initial model.

When evaluating, as expected, the model's accuracy and loss scores on the test set are greater compared to the results during training, with 86.43% accuracy and 33.81% loss score, as we can see in Figure 8.

```
model_cnn.evaluate(test_set)

2/2 [==============================] - 0s 58ms/step - loss: 0.3381 - accuracy: 0.8643
[0.3381214439868927, 0.8642857074737549]
```

Figure 8: Model accuracy and loss using test data

Since we are implementing the CNN model in screening for a medical-related problem, it is recommended to use other evaluation metrics for classification tasks such as recall, precision, specificity, and F1 score. Recall or sensitivity is the true positive rate which measures the proportion of actual positives that are identified correctly by the model (Ahmed et al., 2022). This model has a recall score of 81.25%. The opposite of recall is specificity which is the true negative rate that measures the proportion of the actual negatives that are identified correctly by the model (Ahmed et al., 2022). The specificity score of this model is 90.38%. Precision is the positive predictive value that measures the proportion of positive identifications that are correct, and this model has a precision score of 88.64%. Last, the F1 score is the harmonic mean of precision and recall, which measures the preciseness and robustness of the CNN model, and it has an F1 score of 84.78.

Figure 9 shows the results of the mentioned metrics that were manually computed based on the prediction results.

```
Recall: 81.25
Precision: 88.64
Specificity: 90.38
F1 Score: 84.78
```

Figure 9: Initial model evaluation scores in percentage

Amazon SageMaker Model

We measured the same evaluation metrics on the tuned model in Amazon SageMaker. The best-tuned model's evaluation scores surpassed the initial model's scores as we can see in Figure 10. The tuned model has an accuracy of 95%, a recall score of 90%, a specificity score of 100%, a precision score of 100%, and an F1 score of 94.74%.

```
accuracy: 95.00
Recall: 90.00
Precision: 100.00
Specificity: 100.00
F1 Score: 94.74
```

Figure 10: Amazon SageMaker best model evaluation scores in percentage

Appendix F shows the predictions of the best-tuned model in SageMaker while Appendix G shows its hyperparameters based on the tuning job results.

Appendix H contains all the results of the tuning jobs in Amazon SageMaker. The first row of the data frame shows the best-tuned model's validation accuracy, which is the Final Objective Value, is 79.07% and is higher than the validation accuracy of the initial model. As expected, the compute time is shorter; it only took 388 seconds (about 6 and a half minutes) to retrain and tune this model.

iOS Application

Figure 11 is a sample output when using the EZ Autism Screener application in an actual iPhone device. The image used is from the Kaggle ASD Facial Images test subset. The iPhone camera was used to capture this image and after verifying of using this image, the application provided a result indicating "Autistic" as we can see on the top portion of the phone screen.

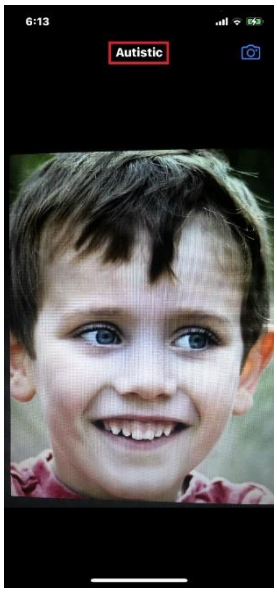


Figure 11: EZ Autism Screener iOS app result

8. CONCLUSION

Even without hyperparameter tuning, the initial model provided very promising evaluation results, most importantly the specificity score, which is an extremely critical metric in medical-related problems, like identifying autism in children. Looking at the specificity score, the model correctly identified 90.38% of those without autism as non-autistic. When diagnosing diseases or disorders, we must ensure that those predicted as negative are negative which is reflected in the specificity score. In addition, the recall score of 81.25% indicates the model correctly identified 81.25% of those with autism as autistic. The precision score of 88.64% means there is an 88.64% probability that those identified as autistic by the model have autism.

The output from Amazon SageMaker retraining and hyperparameter tuning job has exceeded our expectations and produced a superior model at a shorter compute time required for training and tuning the model, in this case, 6 minutes. Based on the accuracy score, the model correctly classified 95% of the images. With the best model's specificity score, the model correctly identified 100% of those without autism as non-autistic, thus ensuring that those predicted as negative are negative. In addition, the recall score of 90% indicates the model correctly identified 90% of those with autism as autistic. The precision score of 100% means there is a 100% probability that those identified as autistic by the model have autism. The model is also very robust and precise as the F1 score of 94.74% indicates.

A great benefit of using Amazon SageMaker is that it reduces the training time from weeks (traditional environment) to hours especially when using the correct instance type for training and tuning models. This saves hours of development. In addition, the model with the best tuning job result can be retrained and tuned further with less time and resources required. This is one of the advantages of using Amazon SageMaker.

We can leverage cloud-based machine learning platform and deep learning models with computer vision, specifically CNNs, to build a universally accessible, user-friendly, and inexpensive tool to screen for autism in children. This is done by integrating the superior model produced from Amazon SageMaker hyperparameter tuning to a mobile app. This broadens access to early detection of autism thereby improving chances of early and effective intervention.

9. FUTURE WORK

We need to improve the application to allow for storing images into Amazon S3 to collect more data with users' consent and approval and if permitted by regulations on collecting sensitive data. This will create a quality dataset of facial images to be available and help with future research on the specific problem. In addition to iOS devices, we need to expand the application to Android devices.

10. REFERENCES

Ahmed, Z., Aldhyani, T., & Jadhav, M. (2022). Facial Features Detection System to Identify Children with Autism Spectrum Disorder: Deep Learning Models. *Hindawi Computational and Mathematical Methods in Medicine*. https://doi.org/10.1155/2022/3941049

Aldridge, K., George, I., Cole, K., Austin, J., Takahashi, N., Duan, Y., & Miles, J. (2011). Facial Phenotypes in Subgroups of Prepubertal Boys with Autism Spectrum

Disorders are Correlated with Clinical Phenotypes. *Molecular Autism, 2*(15). https://doi.org/10.1186/2040-2392-2-15

Amazon SageMaker Model Training. (n.d.). AWS. https://aws.amazon.com/sagemaker/train/

Autism. (2022, March 30). World Health Organization (WHO). Retrieved July 30, 2022, from https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders#:~:text=It%20is%20estimated%20that%20worldwide,prevalence%20varies%20substantially%20across%20studies

Autism Spectrum Disorder. (2022, March). National Institute of Mental Health [NIMH]. Retrieved July 6, 2022, from https://www.nimh.nih.gov/health/topics/autism-spectrum-disorders-asd#:~:text=Autism%20spectrum%20disorder%20(ASD)%20is,first%20two%20years%20of%20life

Badzis, M., & Zaini, M. F. (2014). Early Identification and Intervention of Autism Spectrum Disorder Among Young Children. *IIUM Journal of Educational Studies, 2(1), 67-89*. https://doi.org/10.31436/ijes.v2i1.25

Bauer, K., Morin, K. L., Renz, T. E. III, & Zungu, S. (2022). Autism Assessment in Low- and Middle-Income Countries: Feasibility and Usability of Western Tools. *Sage Journals*. https://doi.org/10.1177/10883576211073691

Beary, M., Hadsell, A., Messersmith, R., Hosseini, M. P., & Soltanian-Zadeh, H. (2022). Diagnosis of Autism in Children Using Facial Analysis and Deep Learning. *Frontiers in Computational Neuroscience*. https://doi.org/10.3389/fncom.2021.789998

Durkin, M. S., Elsabbagh, M., Barbaro, J., Gladstone, M., Happe, F., Hoekstra, R. A., Lee, L. C., Rattazzi, A., Stapel-Wax, J., Stone, W. L., Tager-Flusberg, H., Thurm, A., Tomlinson, M., & Shih, A. (2015). Autism screening and diagnosis in low resource settings: Challenges and opportunities to enhance research and services worldwide. *Autism Res. 8*(5). https://pubmed.ncbi.nlm.nih.gov/26437907/

Data & Statistics on Autism Spectrum Disorder. (2022, March 02). Centers for Disease Control and Prevention (CDC). Retrieved July 5, 2022, from https://www.cdc.gov/ncbddd/autism/data.html

Egger, H. L., Dawson, G., Hashemi, J., Carpenter, K. L. H., Espinosa, S., Campbell, K., Brotkin, S., Schaich-Borg, J., Qui, Q., Tepper, M., Baker, J. P., Bloomfield, R. A. J., & Sapiro, G. (2018). Automatic emotion and attention analysis of young children at home: a ResearchKit autism feasibility study. *Npj Digital Med*, *1*(20). https://doi.org/10.1038/s41746-018-0024-6

Mm909. (2020). Kaggle-Autism [Data]. GitHub, https://github.com/mm909/Kaggle-Autism

Hong, D., Zheng, Y. Y., Xin, Y., Sun, L., Yang, H., Lin, M. Y., Liu, C., Li, B. N., Zhang, Z. W., Zhuang, J., Qian, M. Y., & Wang, S. S. (2021). Genetic syndromes screening by facial recognition technology: VGG-16 screening model construction and evaluation. *Orphanet Journal of Rare Diseases*, *16*(344). https://doi.org/10.1186/s13023-021-01979-y

Kandel, I., & Castelli, M. (2020). The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset. *ScienceDirect*. https://doi.org/10.1016/j.icte.2020.04.010

Koegel, L. K., Koegel, R. L., Ashbaugh, K., & Bradshaw, J. (2014, December 11). *The Importance of Early Identification and Intervention for Children with Or at Risk for Autism Spectrum Disorders.* Taylor & Francis Online. Retrieved July 6, 2022, from https://doi.org/10.3109/17549507.2013.861511

Lu, A., & Perkowski, M. (2021). Deep Learning Approach for Screening Autism Spectrum Disorder in Children with Facial Images and Analysis of Ethnoracial Factors in Model Development and Application. *Brain Science*, *11*, 1446. https://doi.org/10.3390/brainsci11111446

Mussar, M. (2020, April 27). Kaggle - Autism [electronic resource: dataset]. https://github.com/mm909/Kaggle-Autism

Rahman, K. K. M., & Subashini, M. M. (2022). Identification of Autism in Children Using Static Facial Features and Deep Neural Networks. *Brain Science*, *12*(94). https://doi.org/10.3390/brainsci12010094

Sewani, H., & Kashef, R. (2020). An Autoencoder-Based Deep Learning Classifier for Efficient Diagnosis of Autism. *Children*, *7*(0182). https://doi.org/10.3390/children7100182

Yin, W., Mostafa, S., & Wu, F. X. (2021). Diagnosis of Autism Spectrum Disorder Based

on Functional Brain Networks with Deep Learning. *Journal of Computational Biology*, *28(2),* 146-165. https://doi.org/10.1089/cmb.2020.0252

APPENDIX A
Synthesis Matrix of Related Work

	Ahmed, et al., (2022)	Beary, et al., (2022)	Egger, et al., (2018)	Lu, A., & Perkowski, M. (2021)	Rahman, K. K. M., & Subashini, M. M. (2022)	Sewani, H., & Kashef, R. (2020)	Yin, W., Mostafa, S., & Wu, F.X. (2021)	Ata, C. (2022)
Dataset	Kaggle ASD Facial Image	Kaggle ASD Facial Image	Videos collected from participants.	Kaggle ASD Facial Image and East Asia ASD Children Facial Image	Kaggle ASD Facial Image	ABIDE	ABIDE	Kaggle ASD Facial Image
Model	Transfer learning: MobileNet, Xception, and Inception V3.	Transfer learning: MobileNet	Autoencoding with computer vision for analysis.	Transfer learning: VGG-16 trained it in 2 different datasets.	Transfer learning: MobileNet, Xception, and EfficientNet.	Hybrid approach: Autoencoder-KNN, Autoencoder-SVM, Autoencoder-Random Forest, & Autoencoder-CNN with k-fold cross-validation	Hybrid approach: DNN trained on the advance features extracted by the AE. DNN with pretrained AE on raw features found in the MRI results.	Simple CNN model with data augmentation and AWS SageMaker hyperparameter tuning
Evaluation	Best model: MobileNet Training: 100% accuracy at 35 epochs Validation: 95% accuracy at 35 epochs	Test: 94.64 accuracy score at around 15 epochs.		***Kaggle dataset:*** 51.3% accuracy, 66.7% F1-score, 75% (African American) and 86.7% (East Asian) FP rates ***East Asian dataset:*** 95% accuracy, 95% F1-score, 6.67% FP rate ***Combined:*** 23.9% (East Asian) FP rates	Best model: Xception Sensitivity - 88.46% Specificity - 91.66% NPV - 88% PPV - 92% AUC - 96.63%	Best performing model: Autoencoder-CNN 84.05% Accuracy 80% Sensitivity 75.3% Specificity	DNN with AE-extracted advance features: 76.2% accuracy 79.7 ROC AUC score. DNN with pre-trained AE on raw features from fMRI 79.2% accuracy and 82.4% ROC AUC score.	No hyperparameter tuning Accuracy: 86.43% Recall: 81.25% Precision: 88.64% Specificity: 90.38% F1 Score: 84.75% With hyperparameter tuning: Accuracy: 95% Recall: 90% Precision: 100% Specificity: 100% F1 Score: 94.74%
Deployment	MobileNet deployment to web app.		iOS Research Kit for iOS used in clinical setting.					iOS mobile app for public use.

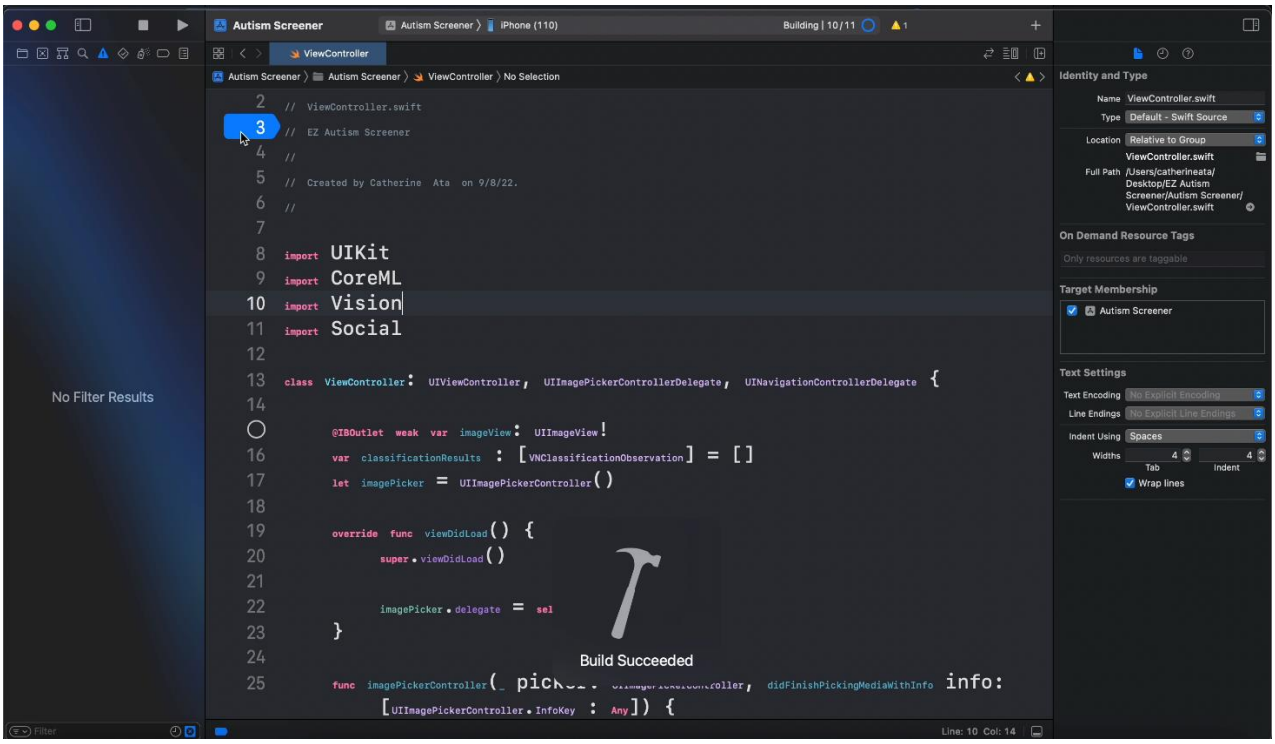
APPENDIX B
Initial CNN Model Summary

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 sequential (Sequential)     (None, 100, 100, 3)       0

 rescaling (Rescaling)       (None, 100, 100, 3)       0

 conv2d (Conv2D)             (None, 100, 100, 32)      416

 max_pooling2d (MaxPooling2D  (None, 50, 50, 32)       0
 )

 conv2d_1 (Conv2D)           (None, 50, 50, 64)        8256

 max_pooling2d_1 (MaxPooling  (None, 25, 25, 64)       0
 2D)

 conv2d_2 (Conv2D)           (None, 25, 25, 128)       32896

 max_pooling2d_2 (MaxPooling  (None, 12, 12, 128)      0
 2D)

 dropout (Dropout)           (None, 12, 12, 128)       0

 conv2d_3 (Conv2D)           (None, 12, 12, 256)       131328

 max_pooling2d_3 (MaxPooling  (None, 6, 6, 256)        0
 2D)

 dropout_1 (Dropout)         (None, 6, 6, 256)         0

 flatten (Flatten)           (None, 9216)              0

 dense (Dense)               (None, 512)               4719104

 dense_1 (Dense)             (None, 2)                 1026

=================================================================
Total params: 4,893,026
Trainable params: 4,893,026
Non-trainable params: 0
```

APPENDIX C
Amazon SageMaker CNN Model Summary

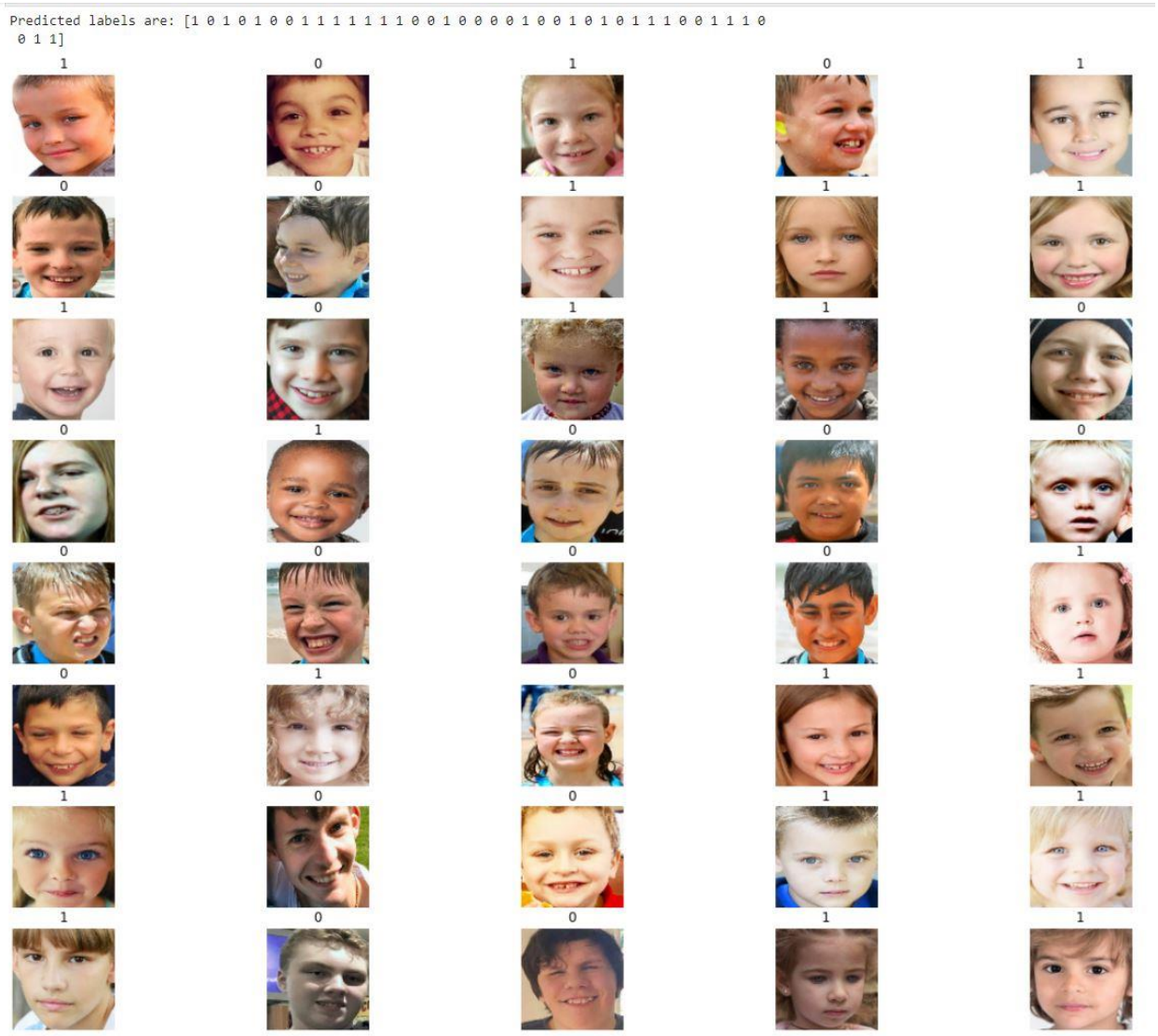
```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 100, 100, 32)      416
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 50, 50, 32)        0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 50, 50, 64)        8256
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 25, 25, 64)        0
_________________________________________________________________
dropout (Dropout)            (None, 25, 25, 64)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 25, 25, 128)       32896
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 12, 12, 128)       0
_________________________________________________________________
dropout_1 (Dropout)          (None, 12, 12, 128)       0
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 12, 12, 256)       131328
_________________________________________________________________
max_pooling2d_3 (MaxPooling2 (None, 6, 6, 256)         0
_________________________________________________________________
dropout_2 (Dropout)          (None, 6, 6, 256)         0
_________________________________________________________________
flatten (Flatten)            (None, 9216)              0
_________________________________________________________________
dense (Dense)                (None, 512)               4719104
_________________________________________________________________
dense_1 (Dense)              (None, 2)                 1026
=================================================================
Total params: 4,893,026
Trainable params: 4,893,026
Non-trainable params: 0
```

APPENDIX D
Building the Application with the Best Autism Model in XCode



APPENDIX E
Initial Model Predictions vs Labels

APPENDIX F
Amazon SageMaker Best-Tuned Model Sample Predictions



APPENDIX G
Hyperparameters of the Best-Tuned Model in Amazon SageMaker

Hyperparameters

Key	Value
_tuning_objective_metric	val_acc
batch-size	32
epochs	50
learning-rate	0.0010000000000000002
model_dir	"s3:/ /tensorflow-training-2022-08-30-10-03-16-889/model"
optimizer	"nag"
sagemaker_container_log_level	20
sagemaker_estimator_class_name	"TensorFlow"
sagemaker_estimator_module	"sagemaker.tensorflow.estimator"
sagemaker_job_name	"tensorflow-training-2022-08-30-10-03-16-889"
sagemaker_program	"train-cnn.py"
sagemaker_region	"us-east-1"
sagemaker_submit_directory	"s3:// /tensorflow-training-2022-08-30-10-03-16-889/source/sourcedir.tar.gz"

APPENDIX H
AmAmazon SageMaker Tuning Job Results

	batch-size	epochs	learning-rate	optimizer	TrainingJobName	TrainingJobStatus	FinalObjectiveValue	TrainingStartTime	TrainingEndTime	TrainingElapsedTimeSeconds
2	32.0	50.0	0.001000	"nag"	tensorflow-training-220830-1003-028-e05ae715	Completed	0.7907	2022-08-30 10:51:47+00:00	2022-08-30 10:58:15+00:00	388.0
3	85.0	25.0	0.001405	"sgd"	tensorflow-training-220830-1003-027-909a63d9	Completed	0.7796	2022-08-30 10:48:43+00:00	2022-08-30 10:51:15+00:00	152.0
1	84.0	19.0	0.001722	"sgd"	tensorflow-training-220830-1003-029-dceb1e33	Completed	0.7796	2022-08-30 10:51:49+00:00	2022-08-30 10:53:56+00:00	127.0
21	32.0	28.0	0.001000	"sgd"	tensorflow-training-220830-1003-009-1f348e2b	Completed	0.7722	2022-08-30 10:17:23+00:00	2022-08-30 10:21:10+00:00	227.0
13	37.0	23.0	0.001122	"sgd"	tensorflow-training-220830-1003-017-8f73bbbb	Completed	0.7704	2022-08-30 10:32:26+00:00	2022-08-30 10:35:44+00:00	198.0
18	32.0	16.0	0.001263	"sgd"	tensorflow-training-220830-1003-012-993df826	Completed	0.7648	2022-08-30 10:23:34+00:00	2022-08-30 10:26:11+00:00	157.0
5	84.0	45.0	0.001403	"sgd"	tensorflow-training-220830-1003-025-3787c11b	Completed	0.7630	2022-08-30 10:44:07+00:00	2022-08-30 10:48:14+00:00	247.0
0	34.0	37.0	0.001016	"adam"	tensorflow-training-220830-1003-030-43d1b4c1	Completed	0.7611	2022-08-30 10:54:25+00:00	2022-08-30 10:59:17+00:00	292.0
22	96.0	42.0	0.001367	"rmsprop"	tensorflow-training-220830-1003-008-208734dc	Completed	0.7611	2022-08-30 10:17:06+00:00	2022-08-30 10:20:53+00:00	227.0
8	100.0	34.0	0.001000	"adam"	tensorflow-training-220830-1003-022-fd6129f5	Completed	0.7593	2022-08-30 10:40:25+00:00	2022-08-30 10:43:32+00:00	187.0
11	46.0	25.0	0.001458	"sgd"	tensorflow-training-220830-1003-019-b3b0c931	Completed	0.7593	2022-08-30 10:35:33+00:00	2022-08-30 10:38:35+00:00	182.0
12	83.0	26.0	0.001154	"sgd"	tensorflow-training-220830-1003-018-e1ba2dbc	Completed	0.7593	2022-08-30 10:32:27+00:00	2022-08-30 10:35:15+00:00	168.0
20	100.0	12.0	0.001000	"sgd"	tensorflow-training-220830-1003-010-ccbb1979	Completed	0.7537	2022-08-30 10:21:40+00:00	2022-08-30 10:23:17+00:00	97.0
17	32.0	13.0	0.001203	"sgd"	tensorflow-training-220830-1003-013-a3dd9f0f	Completed	0.7444	2022-08-30 10:24:43+00:00	2022-08-30 10:26:55+00:00	132.0
19	32.0	19.0	0.001059	"nag"	tensorflow-training-220830-1003-011-a4654592	Completed	0.7426	2022-08-30 10:21:42+00:00	2022-08-30 10:24:30+00:00	168.0
24	86.0	46.0	0.001804	"sgd"	tensorflow-training-220830-1003-006-6541ddeb	Completed	0.7389	2022-08-30 10:13:05+00:00	2022-08-30 10:17:07+00:00	242.0
15	32.0	35.0	0.001470	"sgd"	tensorflow-training-220830-1003-015-816f559e	Completed	0.7315	2022-08-30 10:27:10+00:00	2022-08-30 10:31:47+00:00	277.0
10	44.0	50.0	0.001057	"sgd"	tensorflow-training-220830-1003-020-99c8ea66	Completed	0.7315	2022-08-30 10:35:59+00:00	2022-08-30 10:42:27+00:00	388.0
4	45.0	34.0	0.001293	"sgd"	tensorflow-training-220830-1003-026-92fbe283	Completed	0.7296	2022-08-30 10:46:24+00:00	2022-08-30 10:51:22+00:00	298.0
9	64.0	5.0	0.001389	"sgd"	tensorflow-training-220830-1003-021-f78d51ec	Completed	0.7241	2022-08-30 10:39:02+00:00	2022-08-30 10:40:11+00:00	69.0
14	100.0	19.0	0.001664	"rmsprop"	tensorflow-training-220830-1003-016-f65b970b	Completed	0.7204	2022-08-30 10:29:34+00:00	2022-08-30 10:31:57+00:00	143.0
6	100.0	50.0	0.001557	"rmsprop"	tensorflow-training-220830-1003-024-1d396641	Stopped	0.5093	2022-08-30 10:43:48+00:00	2022-08-30 10:44:55+00:00	67.0
23	96.0	25.0	0.008969	"rmsprop"	tensorflow-training-220830-1003-007-b633d3e9	Completed	0.5093	2022-08-30 10:14:01+00:00	2022-08-30 10:16:39+00:00	158.0
25	96.0	30.0	0.008933	"rmsprop"	tensorflow-training-220830-1003-005-126798da	Completed	0.5093	2022-08-30 10:10:31+00:00	2022-08-30 10:13:28+00:00	177.0
16	32.0	50.0	0.001000	"rmsprop"	tensorflow-training-220830-1003-014-418cc33c	Stopped	0.4907	2022-08-30 10:26:34+00:00	2022-08-30 10:27:57+00:00	83.0
26	66.0	44.0	0.003749	"adam"	tensorflow-training-220830-1003-004-4f193d33	Completed	0.4907	2022-08-30 10:08:42+00:00	2022-08-30 10:12:50+00:00	248.0
27	33.0	17.0	0.008287	"rmsprop"	tensorflow-training-220830-1003-003-5f9c0753	Completed	0.4907	2022-08-30 10:07:38+00:00	2022-08-30 10:10:10+00:00	152.0
28	92.0	33.0	0.006509	"rmsprop"	tensorflow-training-220830-1003-002-8e0680c7	Completed	0.4907	2022-08-30 10:05:02+00:00	2022-08-30 10:08:29+00:00	207.0
29	39.0	10.0	0.003516	"adam"	tensorflow-training-220830-1003-001-c6a3bb88	Completed	0.4907	2022-08-30 10:04:44+00:00	2022-08-30 10:07:13+00:00	149.0
7	49.0	32.0	0.001465	"sgd"	tensorflow-training-220830-1003-023-94c757e5	Stopped	0.4815	2022-08-30 10:43:02+00:00	2022-08-30 10:43:55+00:00	53.0

Are Companies Responsible for Internet of Things (IoT) Data Privacy? A Survey of IoT User Perceptions

Karen Paullet
paullet@rmu.edd
Robert Morris University
Pittsburgh, PA

Adnan A. Chawdhry
chawdhry_a@pennwest.edu
Pennsylvania Western University
California, PA

Jamie Pinchot
pinchot@rmu.edu
Robert Morris University
Pittsburgh, PA

Abstract

This study addressed privacy concerns of Internet of Things (IoT) users, in relation to concerns about personal data collection. Data breaches continue to impact people who use online services such as web sites, mobile apps, and IoT devices. IoT devices, in particular, can often collect data via sensors without the user even being aware of all of the varied types of data being collected. Therefore, this study examined IoT users' data privacy concerns perceptions regarding the responsibilities of companies providing IoT devices and data collection services. A survey of 353 IoT users was conducted and found that participants had a high level of concern for data privacy and a high level of concern about ethical violations at companies that provide and collect data from IoT devices. The survey focused on the users experience across all IoT devices but did have one question to identify IoT devices they have used. The majority of participants had experienced a prior data privacy violation, and prior experience did impact their privacy concerns. However, prior experience did not impact participant's comfort level with allowing data collection, and participants also indicated that the benefits of sharing IoT data could outweigh the data privacy risks.

Keywords: data privacy, privacy, Internet of Things, mobile devices, ethics

1. INTRODUCTION

The Internet of Things (IoT) refers collectively to the many and varied types of devices that can connect to the Internet. These devices are often referred to as "smart" devices and can range from personal devices to smart home appliances and smart city devices. Personal IoT devices can include smart phones, smart watches, health and fitness trackers, and other wearables. Examples of home IoT devices include toasters, refrigerators, thermostats, light switches, video

monitors, and doorbell cameras that can all be controlled via mobile apps or web sites because they are connected to the Internet. At the largest scale, smart city IoT devices can include smart traffic lights that sense and adjust to traffic patterns, surveillance cameras, and trackable bicycles and scooters (Haney et al., 2021; Rice & Bogdanov, 2019; Zheng et al., 2018).

While these smart devices can provide many conveniences for people, they also introduce some new threats in regard to data privacy. While most Internet users are aware of the data they are sharing online via web sites, mobile apps, and social media platforms, the data shared via IoT devices can be less obvious, and many users are not even aware that certain data is being collected. Non-technical users in particular may not understand a device's privacy and security implications (Rice & Bogdanov, 2019). Some examples of data collected by sensors on IoT devices that may impact personal privacy and safety include current location, past locations, and even most frequented locations. When traveling, sensors can determine changes in direction, speed, and acceleration. Health and fitness devices may include sensor data that tracks sensitive health information (Zheng et al., 2018).

Any personal data that is collected and stored can potentially be breached. IoT devices exacerbate the problem of data privacy by generating an exponentially increased amount of personal data, often without the knowledge of the user. Further, IoT devices currently face a number of security challenges and lack of regulation. Due to these issues, IoT devices present a serious threat to data privacy (Cirne et al., 2022; Foltz & Foltz, 2021; Rice & Bogdanov, 2019).

Because of the lack of regulation on IoT devices, and the continual development of new types of devices, it is important for users to be aware of the ways in which IoT data is collected and shared (Rice & Bogdanov, 2019). It is currently unclear where perceived responsibility for IoT data privacy lies.

Haney et al. (2021) conducted a study of smart home users and found that users assume some personal responsibility for data privacy but also assign responsibilities to manufacturers and government. Users may mistakenly believe that IoT manufacturers have taken precautions for data privacy. However, not all companies see online privacy as a corporate social responsibility. Pollach (2011) found that only a small proportion of information technology companies have implemented comprehensive privacy programs. Allen and Peloza (2015) found that despite its importance in a world of digital technologies, the concept of privacy is rarely addressed in research on corporate social responsibility. Further, Rice and Bogdanov (2019) found that methods companies typically use to describe their privacy practices, such as the privacy statement available on product websites, are largely ineffective in conveying information to users.

Because of this potential threat to data privacy that IoT devices present, it is important to understand IoT users' perceptions of data privacy, and their perceptions of the responsibilities of companies that provide IoT devices and collect data from them.

2. LITERATURE REVIEW

Personally Identifiable Information (PII), as defined by NIST (2017) is "information which can be used to distinguish or trace an individual's identity, such as their name, social security number, biometric records, etc. alone, or when combined with other personal or identifying information which is linked or linkable to a specific individual, such as date and place of birth or mother's maiden name." PII is often the target of a data breach. For IoT users, PII could also include sensor data that could impact privacy, such as personal health information. It could also include sensor data that could impact personal safety, such as current location or location history. Because so much data is generated by IoT devices, a data breach could be a serious privacy threat for users.

Recent Data Breaches

Data breaches of sensitive information are on the rise. According to the Identity Theft Resource Centers 2021 Annual Data Breach Report, the overall number of data compromises is up more than 68% compared to 2020. The new record number of data compromises is 23% over the previous all-time high set in 2017. Additionally, the number of data events that involved sensitive information such as social security numbers increased to 83% from 80% in 2020 (ITRC, 2022). Other key findings in the report include:

- Ransomware-related data breaches have doubled in each of the past two years (2020, 2021). At the current rate, ransomware attacks will surpass phishing as the number one root cause of data compromises in 2022.
- There were more cyberattack-related data compromises (1,608) in 2021 than all data compromises in 2020 (1,108).

- Compromises increased from 2020 to 2021 in every primary sector but one, the military, where there were zero public breaches.
- The number of data breach notices that do not reveal the root cause of a compromise has grown by more than 190% since 2020.

Cyberattacks and security breaches continue to happen daily around the world. In the United States in 2022 below are a partial list of breaches:

- Crypto.com (Jan 17): targeted nearly 500 cryptocurrency wallets
- Red Cross (Jan): An attack on a third-party contractor had more than a half a million records compromised
- GiveSendGo (Feb): A political hacker stole then published the information of 90,000 people who donated money to protestors
- Flagstar Bank (June): The Michigan-based bank notified 1.5 million customers that hackers stole their social security numbers
- Marriott: Marriott International confirmed that hackers stole 202 gigabytes of sensitive data

It is important to mention that cyber-attacks worldwide are growing rapidly. Approximately 95% of cybersecurity breaches are caused by human error (World Economic Forum, 2020). The U.S. was the target of 46% of cyberattacks in 2020, more than double any other country (Lambert, 2021). Fifty-four percent of companies say their IT departments are not sophisticated enough to handle advanced cyberattacks (Sophos, 2021). Data breaches exposed 22 billion records in 2021 (Risk Based Security, 2022).

There has been a surge of interconnected devices known as IoT (Internet of Things). With this rapid growth of IoT enabled devices breaches are on the rise. The number of connected IoT devices as of September 2022 is 14.3 billion globally (Hasan, 2022). This is expected to grow to 75 billion by 2025. Approximately 84% of surveyed companies have reported an IoT security breach (Conosco, 2021). Ring, owned by Amazon had two separate incidents where user data was exposed to a third party. One where trackers were embedded into their Android application and the second due to an IoT security breach where cybercriminals hacked into the home monitoring systems of several families

According to Gartner, 40% of smart home appliances globally are being used for botnet attacks (Gartner, 2021). Research given to the FDA found that St. Jude Medical's implantable devices have vulnerabilities. If hackers were able to gain access they could deplete the battery or administer incorrect pacing shocks.

Consumer Privacy Concerns

SAS (2018) found that although consumers acknowledge their own responsibility for their personal data, 73% or participants believe that organizations are collecting their personal information without their knowledge. Fifty-eight percent of respondents said they do not trust organizations to keep their personal information secure while believing that 57% of companies do not try their best to protect consumers data. When it comes to industries that people trust most to protect their data, 46.5% of participants believe that health care and banking are the most secure. Social media was the least trusted, with only 14% of participants expressing the same confidence, followed by retail at 18%, energy companies at 21%, and government agencies at 29%.

Consumers' stated privacy preferences, as measured in surveys, can often differ from their actual behavior, as measured by consumers' online activity. This is referred to as the "privacy paradox". Research on the privacy paradox explains that consumers may judge privacy as important in surveys but continue to engage with websites and disclose information (Martin, 2019; Kokolakis, 2017; Strahilevitz et al., 2016; Norberg et al., 2007). The privacy paradox is important for businesses because the narrative defines the scope for corporate responsibility as quite narrow. Companies have little to no responsibility to identify or respect privacy expectations of consumers while online (Martin, 2019). So, in practice, consumers essentially must give up their right to privacy when they go online, use social media, or use a mobile app.

The privacy paradox suggests that consumers demonstrate their willingness to 'trade' the risk of privacy for the benefits of sharing information online. Consumers regularly exchange their privacy preferences for the benefits of discounts, better service, or social affiliation (Martin, 2019; Schumann et al., 2014; Xu et al., 2009; Hui et al., 2007). This exchange approach to privacy shows consumers as taking the risks and benefits of disclosing information into consideration when assessing privacy concerns. Consumers are willing to disclose for personalization and free services (Martin, 2019; Xu et al., 2009; Banerjee et al., 2008).

Consumer concerns about the security of their data continues to solidify as cyber-attacks

continue to grow. Companies' information security practices are increasingly the subject of government scrutiny through the Federal Trade Commission (FTC), the Financial Industry Regulatory Authority (FINRA) and the Health Insurance Portability and Accountability Act (HIPPA). Additionally, the Securities and Exchanges commission (SEC) has elevated the issue of cybersecurity to the level of the board of directors of public companies (Aguilar, 2014).

Consumer Attitudes Toward Data Breaches

Mayer et al. (2021) conducted a study on individuals' awareness, perceptions and responses to data breaches. The study found that 73% of participants experienced at least one breach and 5.36 breaches on average. An email address's likelihood of being exposed in a breach significantly correlated with the email account's age and utilization. Only 14% of participants attributed the cause of being affected by a breach to external factors such as hacking. Most participants rated their concern regarding breaches as low (56% slightly/somewhat concerned, 19% no concern). Breaches such as the release of their physical address or passwords raised more concern. Lastly, participants reported having already changed or being very likely to change their passwords and review their credit report and financial statements in response to over 50% of breaches.

A 2019 study conducted by the Pew Research Center in regard to privacy and personal data revealed that seven out of 10 Americans feel as if their data is less secure than it was five years prior to the study date. Roughly three out of 10 Americans have experienced some kind of data breach in the previous 12-month period and eight out of 10 believe they have control over their personal data. Lastly, only 6% of adults say they understand what companies do with the data collected (Pew Research Center, 2019).

3. PURPOSE

The purpose of this study was to explore the perceptions of users of Internet-connected devices in regard to data privacy responsibilities of the companies that collect data from these devices. Further, the study examines whether prior experience with a data privacy issue, such as involvement in a data breach, impacts those perceptions. The following research questions were addressed in the study:

RQ1: What are users' perceptions about the responsibilities of companies collecting user data from Internet-connected devices in regard to data privacy?

RQ2: How does prior experience with a data privacy issue impact users' data privacy concerns?

4. METHOD

This study used a survey research method (Fowler, 2013) and collected data via an electronic survey. The population for the study consisted of adults aged 18 and older who own and have used at least one Internet-connected device. There were 353 responses collected (n=353). The study was approved by the university's Institutional Review board (IRB).

The survey included questions addressing general demographic data: age group, gender, and whether the participant works in a technology-related field. The next set of questions addressed users' comfort level with companies collecting personal data from Internet-connected devices. Another set of questions addressed user concerns about whether allowing devices to collect personal data could lead to ethical violations at the company that collects the data. Questions specifically addressed whether users believe it is the responsibility of the company to protect the privacy of data collected from personal devices, and further asked about what specific responsibilities should be upheld, if any, on the part of the company collecting data. Participants were asked if they think that the safeguards put in place by companies to protect personal data privacy are adequate. Finally, participants were asked whether they had experienced any prior issues with data privacy related to Internet-connected devices (adapted from Xu et al., 2012), about the personal impact, if any, of allowing devices to collect their personal data, and whether they planned to change any behavior in regard to sharing data from devices in the future.

Measuring Privacy Concern

The Mobile Users' Information Privacy Scale (MUIPC) was used to measure the participants' privacy concerns regarding Internet-connected devices. MUIPC was developed by Xu et al. (2012) and was partially based on both the Concern for Information Privacy (CFIP) scale (Smith et al., 1996) and the Internet User's Information Privacy (IUIPC) scale (Malhotra et al., 2004). Malhotra et al. (2004) adapted CFIP for the online environment in developing IUIPC (Malhotra et al., 2004; Smith et al., 1996; Xu et al., 2012). MUIPC further developed the

questions to apply to mobile device and app users in regard to data privacy (Xu et al., 2012). MUIPC has been used to address Internet-connected devices, commonly referred to as the Internet of Things, as they fall into the category of mobile devices (Foltz & Foltz, 2020; Foltz & Foltz, 2021; Pinchot & Cellante, 2021). This makes the use of MUIPC appropriate for this study.

MUIPC is a scale consisting of 9-items, each of which is measured on a five-point Likert scale ranging from "Strongly Disagree" = 1 to "Strongly Agree" = 5. The scale measures three dimensions of mobile data privacy concern: perceived surveillance, perceived intrusion, and secondary use of personal information (Xu et al., 2012). Surveillance includes any collection or processing of personal data in order to influence the individuals from whom the data has been collected (Lyon, 2001). Methods of data collection can include watching, listening to, or recording individuals' actions or conversations (Solove, 2006). Perceived intrusion is having more personal information shared about oneself than an individual is comfortable with having shared (Xu et al., 2012). Finally, secondary use of information refers to the concern that personal data will be used without permission in an undisclosed or unexpected way (Smith et al., 1996; Xu et al., 2012). The MUIPC scale has been tested for internal consistency, with a Cronbach alpha coefficient above .7 (Xu et al., 2012; Degirmenci et al., 2013), which is a high score.

Sample

This study utilized Amazon Mechanical Turk (MTurk) for sample selection and distribution of the electronic survey. MTurk is a crowdsourcing tool that allows access to participants that meet specific inclusion criteria who are willing to participate in surveys for compensation. MTurk has been found to be largely representative of the entire U.S. population, and is used widely in academic research (Lovett, 2018; Redmiles et al., 2019). In MTurk, compensation is offered to all participants; the researcher chooses the amount of compensation to offer, and the number of respondents desired. For short surveys (approximately 5-9 minutes), the compensation amount offered is typically between $.10 and $.50 (Lovett, 2018). This survey had an average completion time of 6 minutes. Compensation was provided within the recommended range.

Question Pro was used to create the electronic survey and record responses. The electronic survey was posted on MTurk in May 2022 targeting 350 responses. A total of 384 people began the survey, and 353 people submitted complete surveys (n=353). The high response rate, 92%, is typical of using the MTurk platform.

5. RESULTS

The survey began with questions to evaluate background demographics. Of the participants, half were within the 25-34 age group while nearly a quarter of the respondents were within the 35-44 age group. Table 1 provides the frequency distribution of participants' ages. Additionally, the gender breakdown of the participants was 40.06% female and 59.94% male. The majority of the participants, 91.2%, stated they work in a technology-related field while 8.38% stated they do not.

Age Group	Percentage
18-24	4.00%
25-34	50.00%
35-44	24.29%
45-54	16.29%
55-64	5.14%
Above 64	0.29%

Table 1: Age Distribution

Addressing RQ1

The first research question focused on the users' perceptions on a company's responsibility when collecting user data. Of the participants, 91.6% stated they have a concern that companies collecting data from devices can lead to an ethical violation. A breakdown of their responses can be found in Table 2. Additionally, 95.3% of the participants felt that it is the company's responsibility to protect the data it collects, while 4.7% did not. Subsequent to this question, participants were asked if the safeguards put in place by organizations were adequate and only 85.5% stated yes while 14.5% stated no, they are not adequate.

Concern for Company Ethical Violations	Response
Yes - I have already experienced an ethical violation related to collection of my data	68.1%
Yes - I am concerned that I will experience an ethical violation related to collection of my data	23.5%
No	8.4%
Total	100%

Table 2: Ethical Violations

Participants were asked an open-ended question about the ethical responsibilities of companies when collecting user data. A few notable statements written by participants are:

- *Now more than ever, you should know how big data is collected and understand some of the impacts of big data in your personal life.*
- *Big data helps us save money on what we eat too with loyalty card schemes, cashback sites and money-off coupons all designed to reduce the weekly food bill. . A staggering 90% of the world's data has been created in the past two years.*
- *Improving health care and generating scientific knowledge create an ethical imperative for the sharing of data. Sharing data, if done appropriately, can help to address health inequalities, and therefore creates an obligation to participants who have consented to use the data well and efficiently.*
- *Smart business leaders and key stakeholders are making it a priority to implement corporate responsibility programs (CSR). A CSR program can support worthy causes, improve employee morale, and create a company culture of integrity.*
- *The ethical responsibilities that companies have to customers revolve around collecting only necessary data from customers properly protecting customer data.*

Addressing RQ2

The second research question evaluates users' prior experiences with data privacy issues and whether those experiences impact their data privacy concerns.

To measure user's privacy concerns, the MUIPC scale was used to create an index PRIVACY CONCERN variable. The score for PRIVACY CONCERN could range from a minimum of 3 to a maximum of 45. Since the scale questions used a 5-point Likert scale scored from "Strongly Agree" = 1 to "Strongly Disagree" = 5, the scale was inverted, with a low score indicating a high level of concern, and a high score indicating a low level of concern. The scores were then categorized as either High Privacy Concern (24 or less) or Low Privacy Concern (25 to 45).

Scores ranged from 6 to 39 with 87.22% in the High Privacy Concern category and 12.78% in the Low Privacy Concern category. This shows that there was clearly a high level of concern in regard to data privacy for this sample. The scale showed good internal consistency (Cronbach's a=.84).

The participants were asked about their overall impact if they allowed devices to collect data about them. Of the participants, 83.4% stated it had some impact (positive or negative) while 13.1% stated it had no impact on their lives. The breakdown of responses is available in Table 3.

Collecting Data Impact	Percent
It had a positive impact on my life	53.80%
It had a negative impact on my life	29.60%
It had no impact on my life	13.10%
I do not allow data collection on my devices.	3.40%
Total	100%

Table 3: Impact of Prior Data Collection

To get a better sense of the participants privacy concerns, the researchers tested two variables (PRIOR EXPERIENCE and ETHICAL VIOLATIONS CONCERN) against both PRIVACY CONCERN and COMFORT WITH DATA COLLECTION. Statistical significance (p-value of less than or equal to 0.05) was found in all cases. The results of this analysis can be found in Tables 5 and 6.

Variable	Chi-square Value	df	p-value (* indicates statistical significance)
Prior Privacy Violations	405.21	6	.000*
Ethical Violations Concern	13.421	2	.001*

Table 4: Privacy Concern Chi-Square

Table 4 shows that there is a strong statistically significant relationship (p<.000) between PRIOR PRIVACY VIOLATIONS and PRIVACY CONCERN. This indicates that the more prior experiences a participant had with privacy violations such as data breaches or ethical violations, the higher their level of privacy concern.

Additionally, there is a strong statistically significant relationship (p<.001) between ETHICAL VIOLATIONS CONCERN and PRIVACY CONCERN. This indicates that the more concern a participant has that companies will allow ethical violations with their data, the higher their level of privacy concern.

Table 5 shows that there is a statistically significant relationship (p<.029) between PRIOR PRIVACY VIOLATIONS and COMFORT WITH DATA

COLLECTION. This indicates that the more prior experiences a participant had with privacy violations such as data breaches or ethical violations, the higher their level of comfort with data collection. This result is counter-intuitive, but may indicate that users who have already experienced data privacy violations are no longer as concerned about them**.**

Variable	Chi-square Value	df	p-value (* indicates statistical significance)
Prior Privacy Violations	7.093	2	.029*
Ethical Violations Concern	35.516	2	.000*

Table 5: Comfort with Data Collection Chi-Square

Additionally, there is a strong statistically significant relationship (p<.000) between ETHICAL VIOLATIONS CONCERN and COMFORT WITH DATA COLLECTION. This indicates that the more concern a participant has that companies will allow ethical violations with their data, the higher their level of comfort with data collection. This result is also counter-intuitive, but may indicate that while participants are concerned with potential ethical violations, they find that the benefits outweigh the risks.

6. DISCUSSION

The primary focus of this study was to explore the perceptions of users of Internet-connected devices in regard to data privacy responsibilities of the companies that collect data from these devices. User perceptions were assessed by questions related to concerns about companies having an ethical violation with data collected, the company's responsibility to protect user privacy, and the adequacy of a company's safeguards. It was interesting to note that 91.6% of the participants had some level of concern and 68.1% of the participants actually had experienced some kind of ethical violation. Of the participants, 95.3% responded that a company is responsible for protecting data collected on their devices, which seems in line with the expectations. Of the participants, 85.3% stated that the companies had put in proper safeguards to protect their privacy. Considering these responses, one theory is that users feel that companies are safeguarding their data but are nevertheless still concerned about potential ethical violations.

After reviewing the open-ended question about the users' perceptions on a company's ethical responsibilities, it was clear that users felt strongly that a company needs to go above and beyond to protect data privacy if the need should arise to collect data. Several responses noted that the benefits that could be obtained with collecting this data could outweigh the impact of an ethical violations. This is a valuable finding because it shows that some users find the benefits providing by data collection to outweigh the risks. Most important was to see that users appreciate when companies' setup internal departments to help ensure that consumer data is protected to avoid an ethical violation. Lastly, the researchers found it interesting that a few participants mentioned that in the time of internet-connected devices, it is also our responsibility to understand what data is collected, how it is used, and how big data can be impactful in our lives. This comment was important because it showed that not only is the responsibility on the company, but some responsibility should be shared with the individual using these devices.

The second research question focused on the participants' prior experiences with data collection from devices and the impact it has had on their life. A very low percentage stated they either had no impact in their lives or do not allow companies to collect data. Most interesting was that 53.8% of the participants had a positive impact from the data being collected while 29.6% did not. The split between these two impacts seems proportionate to the open-ended comments as some participants stated a strong desire for additional protection and others were accepting of data collection but with a beneficial trade-off like improved healthcare or user experience.

The research also found that participants had a higher comfort level with data collection when they had less prior experience with a privacy issue such as a data breach or ethical violation. An indirect relationship was found between comfort collecting data and the impact of allowing data to be collected. Therefore, the higher comfort level was linked with a positive impact while a lower comfort level was highest with a negative impact. While these two relationships were indirect, they do add to the findings of the study that comfort allowing data collection does have a statistically significant relationship with their prior experience of a privacy issue such as a data breach or ethical violation.

7. CONCLUSIONS

It is clear that there are serious data privacy concerns for users of IoT devices (Cirne et al., 2022; Foltz & Foltz, 2021; Rice & Bogdanov, 2019) and for companies that provide these devices (Martin, 2019; Acquisti et al., 2006).

This study explored the perceptions of users of Internet-connected (IoT) devices in regard to data privacy responsibilities of the companies that that provide the devices and collect data from them. A clear majority of participants, 95.3%, responded that a company is responsible for protecting data collected on their devices. Further, 85.3% felt that companies did provide adequate safeguards but 91.6% stated that they were concerned about potential ethical violations by companies collecting their data. Collectively, these findings indicate that there is a high level of concern for data privacy, but participants also felt that companies are implementing adequate safeguards and the benefits could outweigh the risks when it comes to allowing IoT data collection.

Additionally, the study examined whether prior experiences with a data privacy violation such as a data breach or ethical violation impacted an IoT user's level of privacy concern. In the sample, a majority, 68.1% had experienced a prior privacy violation. Privacy concern was measured using the MUIPC scale (Xu et al., 2012). The majority of participants, 87.22%, fell into the High Privacy Concern category, indicating that privacy is a concern for the participants. Statistically significant relationships were found between privacy concern and prior privacy violations, and between privacy concern and ethical violations concern. These findings clearly show that high levels of privacy concern are related to past experience with data breaches or ethical violations and specific concern for how companies handle personal data.

More surprising were the statistically significant relationships found between comfort level in allowing data collection and both prior privacy violations and ethical violations concern. These findings seem counter-intuitive and were unexpected. They indicate that as the IoT users' prior experiences with data breaches or other privacy violations increases, their comfort level with data collection increases. Similarly, as the IoT users' concern for ethical violations by companies increases, their comfort level with data collection increases.

A potential limitation of this study was the use of Amazon Mechanical Turk to recruit participants.

MTurk has been known to skew toward tech-savvy participants even though it has been shown to be representative of the overall U.S. population (Lovett, 2018; Redmiles et al., 2019. For this sample in particular, 91.62% of participants reported that they work in a technology-related field. This type of work could potentially mean that users in this sample are more aware of the potential risks to data privacy when using IoT devices that other non-technical users.

This exploratory study included some contradictory findings and future studies could further examine the complex perceptions of IoT users about data privacy. Future studies could also focus on corporate responsibilities and address a population of IoT companies to explore the corporate perspective on this issue.

8. REFERENCES

Ablon, L., Heaton, P., Lavery, D.C., & Romanosky, S. (2016). Consumer attitudes toward data breach notifications and loss of personal information. Technical report. Rand Corp.

Aguilar, L. (2014). *Boards of directors, corporate governance and cyber-risks: Sharpening the focus.*http://www.sec.gov/News/Speech/Detail/Speech/1370542057946

Allen, A.M., & Peloza, J. (2015). Someone to watch over me: The integration of privacy and corporate social responsibility. *Business Horizons, 58*, 635-642.

Banerjee, S. S. & Dholakia, R.R. (2008). Mobile advertising: Does location based advertising work? *International Journal of Mobile Marketing, 2*(2), 68-74.

Cirne, A., Sousa, P.R., Resende, J.S., & Antunes, L. (2022). IoT security certifications: Challenges and potential approaches. *Computers & Security, 116*, 1-28.

Conosco. (2021). IoT security breaches: 4 Real-world examples. Retrieved from https://www.conosco.com/blog/iot-security-breaches-4-real-world-examples/

Degirmenci, K., Guhr, N., & Breitner, M. (2013). Mobile applications and access to personal information: A discussion of user's privacy concerns. *Proceedings of the 34th International Conference on Information Systems*, 1-21.

Foltz, C.B., & Foltz, L. (2020). Mobile user's

information privacy concerns instrument and IoT. *Information & Computer Security, 28*(3), 359-371.

Foltz, C.B., & Foltz, L. (2021). MUIPC and intent to change IoT privacy settings. *The Journal of Computing Sciences in Colleges, 36*(7), 27-38.

Fowler, F.J. (2013). *Survey research methods (5th edition).* Sage.

Haney, J., Acar, Y., & Furman, S. (2021). It's the company, the government, you, and I: User perceptions of responsibility for smart home privacy and security. *Proceedings of the 30th USENIX Security Symposium*, 411-428.

Hassan, M. (2022). State of IoT 2022: Number of connected IoT devices growing 18% to 14.4. billion globally. IOT Analytics.

Hui, K., Hock, H.T., Sang-Yong, T.L. (2007). The value of privacy assurance: An exploratory field experiment. *MIS Quarterly, 31*(1), 19-33.

Identity Theft Resource Center (2022). *Identity Theft Resource Center's 2021 annual data breach report sets new record for number of compromises.* https://www.idtheftcenter.org/post/identity-theft-resource-center-2021-annual-data-breach-report-sets-new-record-for-number-of-compromises/

Karunakaran, S., Thomas, K., Bursztein, E., & Comanescu, O. (2018). Data breaches: User comprehension, expectations, and concerns with handling exposed data. *Symp. On Usable Privacy and Security*, 217-234.

Kokolakis, S. (2017). Privacy attitudes and privacy behavior: A review of current research on the privacy paradox phenomenon. *Computers & Security, 64*, 122-134.

Lambert, J. (2021). *Microsoft digital defense report shares new insights on nation-state attacks.* Microsoft Security. https://www.microsoft.com/security/blog/2021/10/25/microsoft-digital-defense-report-shares-new-insights-on-nation-state-attacks/

Lovett, M., Bajaba, S., Lovett, M., & Simmering, M. (2018). Data quality from crowdsourced surveys: A mixed method inquiry into perceptions of Amazon's Mechanical Turk Masters. *Applied Psychology, 67*(2), 339-366.

Lyon, D. (2001). Surveillance society: *Monitoring everyday life.* Open University Press.

Malhotra, N.K., Kim, S.S., & Agarwal, J. (2004). Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information Systems Research, 15*(4), 336-355.

Marti, K. (2019). Breaking the privacy paradox: The value of privacy and associated duty of firms. *Business Ethics Quarterly, 1052*, 150.

Mayer, P., Zou, Y., Schaub, F., & Aviv, A. (2021). Now I'm a bit angry: Individuals' awareness, perception, and responses, to data breaches that affected them. *USENIX Security Symposium 2021.*

Mikhed, V., & Vogan, M. (2018). How data breaches affect consumer credit. *Journal of Banking and Finance*, *88*, 192-207.

Norberg, P.A., Horne, D.R., & Horne, D.A. (2007). The privacy paradox? Personal information disclosure intentions versus behaviors. *Journal of Consumer Affairs, 41*(1), 100-126.

NIST. (2017). Information technology laboratory: Computer resource center. https://csrc.nist.gov/glossary/term/personally_identifiable_information

Pew Research Center (2019, November). *Americans and Privacy: Concerned, confused and feeling lack of control over their personal information.* https://www.pewresearch.org/internet/2019/11/15/americans-and-privacy-concerned-confused-and-feeling-lack-of-control-over-their-personal-information/

Pinchot, J., & Cellante, D. (2021). Privacy concerns and data sharing habits of personal fitness information collected via activity trackers. *Journal of Information Systems Applied Research, 14*(2), 4-13.

Pollach, I. (2011). Online privacy as a corporate social responsibility: An empirical study. *Business Ethics: A European Review, 20*(1), 88-102.

Redmiles, E.M., Kross, S., & Mazurek, M.L. (2019). How well do my results generalize? Comparing security and privacy survey results from MTurk, web, and telephone samples. *2019 IEEE Symposium on Security and Privacy*, 1326-1343.

Rice, M.D., & Bogdanov, E. (2019). Privacy in doubt: An empirical investigation of Canadians' knowledge of corporate data collection and usage practices. *Canadian Journal of Administrative Sciences, 36*, 163-176.

Risk Based Security (2022, February). *Data breach report: 2021 year end.* https://www.riskbasedsecurity.com/2022/02/04/data-breach-report-2021-year-end/

SAS. (2018). *Data Privacy: Are you concerned? Insights from a survey of US consumers.* https://www.sas.com/content/dam/SAS/documents/marketing-whitepapers-ebooks/sas-whitepapers/en/data-privacy-110027.pdf

Schumann, J.H., Von Wagenheim, F., & Groen, N. (2014). Targeted online advertising: Using reciprocity appeals to increase acceptance among users of free web services. *Journal of Marketing, 78*(1), 59-75.

Smith, H.J., Milberg, J.S., & Burke, J.S. (1996). Information privacy: Measuring individual's concerns about organizational practices. *MIS Quarterly, 20*(2), 167-196.

Sophos. (2021). *Ransomware recovery cost reaches nearly $2 million, more than doubling in a year: Sophos survey shows.* https://www.sophos.com/en-us/press-office/press-releases/2021/04/ransomware-recovery-cost-reaches-nearly-dollar-2-million-more-than-doubling-in-a-year

Strahilevitz, L.J., & Kugler, M.B. (2016). Is privacy policy language irrelevant to consumers? *The Journal of Legal Studies, 45*(S2), S69-95.

The Federal Trade Commission. (2020). *When information is lost or exposed 2020.* https://www.identitytheft.gov/databreach

Wagenseil, P. (2019). *What to do after a data breach.* https://www.tomsguide.com/us/data-breach-to-dos,news-18007.html

World Economic Forum (2020, December). *After reading, writing and arithmetic, the 4th 'r' of literacy is cyber-risk.* https://www.weforum.org/agenda/2020/12/cyber-risk-cyber-security-education

Xu, H., Rossen, M.B., Gupta, S., & Carroll, J.M. (2012). Measuring mobile user's concerns for information privacy. *Thirty Third International Conference on Information Systems*, 1-16.

Xu, H., Zhang, C., Shi, P., & Song, P. (2009). Exploring the role of overt vs. covert personalization strategy in privacy calculus. *Academy of Management Annual Meeting Proceedings*, *2009*(1), 1-6.

Zheng, S., Apthorpe, N., Chetty, M., & Feamster, N. (2018). User perceptions of smart home IoT privacy. *Proceedings of the ACM on Human-Computer Interaction, 2*, 1-20.

Zou, Y., Roundy, K., Tamersoy, A., Shintre, S., Roturier. J., & Schaub, F. (2020). Examining the adoption and abandonment of security, privacy, and identity theft protection practices. *ACM CHI Conference on Human Factors in Computing Systems.*

E-Commerce Drone Delivery Acceptance: A Study of Gen Z's Switching Intention

Jeffrey P. Kaleta
kaletajp@appstate.edu

Wei Xie
xiew1@appstate.edu

Charlie Chen
chench@appstate.edu

Computer Information Systems
Appalachian State University
Boone, NC 28608

Abstract

E-commerce retailers seek to use drone delivery services as an innovative last-mile delivery option. In addition to the challenges of implementing such innovative technologies, the mechanisms that influence Gen Zers who are digital natives to adopt such technologies is limited but necessary to understand to gain success in the early adoption of drone delivery services. Using the theoretical foundations of innovation diffusion and imitation theories, this study examines the mechanisms that influence an online consumer to switch from using a conventional truck delivery service to a less proven drone delivery service. Within this context we construct a psychometric-based research survey to collect our data and structural equation modeling is used for analysis. The findings suggest that the speed and compatibility of the drone delivery and 'the prevalent herding behavior among Gen Zers are significant predictors of their switching intention. Theoretical and practical implications are shared based on the findings of the study.

Keywords: drone delivery, diffusion of innovation, herd behavior, switching behavior, gen-z

1. INTRODUCTION

An Unmanned Aerial Vehicle (UAV), commonly known as a drone, is an autonomous aircraft without any human onboard (Austin, 2010). Originally developed for military applications, drone technology in the civilian domain has been quickly adopted by various industries such as agriculture (Mogili & Deepak, 2018), disaster management (Tanzi et al., 2016), and healthcare (Yaprak et al., 2021). Drones also offer e-commerce industries a promising solution to the challenges associated with truck-dominated last-mile product delivery (Zhu et al., 2020; Leon et al., 2021) by offering faster delivery time, lower maintenance costs, and environmental friendliness (Lee et al., 2016). With this potential, large domestic retail companies such as Amazon, Google, and Walmart also have started pilot-testing drone delivery services for their needs (Wu & Lin, 2018). Australians have taken the lead in receiving the most products via drone delivery, experiencing a 500% increase (Business Wire, 2021). Globally the use of drone package delivery has grown from $0.68 billion in 2020 to $0.99 billion in 2021, at a compound annual growth rate of 45.5% (Business Wire, 2021). Furthermore, the growth of drone usage is expected to increase

more in the advancing days. The number of commercial unmanned aircraft systems registered with the Federal Aviation Administration that deliver purchased products or goods directly to consumers will rise to 70,000 in 2023, compared to 24,000 in 2020 (eMarketer, 2021).

Inflation, the Ukraine war, and the global pandemic have worsened environmental uncertainty, disrupting global supply chains. Due to the increasing number of unpredictable events, retailers are having difficulty filling and delivering orders on time. Order uncertainty has caused customers to feel anxious and dissatisfied (Yaprak, Kılıç, & Okumuş, 2021). Because Gen Zers are active users of e-commerce, they could feel the most impact during uncertain times. Eighty-seven percent of Gen Zers are Amazon Prime members or had been (Statista, 2022). Gen Zers are time- and cost-sensitive when it comes to online delivery. A fast and cheaper delivery service such as drone delivery could be crucial to Gen Zers during these uncertain times.

As an emerging technology, drone delivery research is limited and nascent, especially in understanding what would impact the acceptance of the last-mile drone delivery service (Yoo et al., 2018) for Gen Zers. Looking through the lens of diffusion of innovation theory (Rogers, 1983) and a herd behavior perspective, this study asks, "*As an emerging service with uncertainty, will drone delivery services offered by online retailers influence the behavior of Gen Zers who are online shoppers to switch to drone delivery?*" Additionally, this paper asks, *"Are there intention differences among different demographic groups of gender, neighborhood style, and Amazon membership?"* The objective is to examine the antecedents of Gen Zers' switching intentions from standard truck delivery to drone delivery and inform e-commerce vendors about developing and promoting this emerging service.

This paper is organized as follows. First, we present a literature summary on innovation diffusion, switching behaviors, and the Gen Z demographic. We then present our hypotheses and our methodology to test those hypotheses. Finally, we close with a summary of implications and future research.

2. LITERATURE REVIEW

Diffusion of Innovation Theory

There is a plethora of research using the Diffusion of Innovation Theory (Rogers, 1995) to help explain the factors that influence people's attitudes toward new innovations (Al-Jabri & Sohail, 2012; Tan & Teo, 2000; Thong, 1999). Since its introduction, this theory has made it to the top list of most popular to study the factors that affect an individual to adopt innovative products, services, or technologies (Al-Jabri & Sohail, 2012). Many studies have applied the theory to different innovative service contexts to study people's adoption intentions and behaviors, including internet banking and mobile applications (Tan & Teo, 2000). Rogers (1995) proposed five constructs in the diffusion of innovations theory (DoI) on the work of Taylor and Todd (1995b) regarding the different dimensions of attitudinal beliefs toward innovations. These five dimensions are relative advantage, compatibility, complexity, trialability, and observability. Relative advantage refers to the extent to which an innovation is perceived to improve upon and/or supersede the performance of prior innovations (Rogers, 1995). The compatibility of innovation describes how familiar it is to existing consumers, consistent with the values of those who intend to adopt the innovation (Rogers, 1995). The degree to which an innovation is difficult to understand or use describes its complexity (Rogers, 1995). How much a person can experiment and use a technology describes its trialability, offering a limited installment of how the innovation functions (Rogers, 1995). Lastly, observability refers to how visible the use and practice of the innovation are in sight and noticeable to others in a social system (Rogers, 1995).

Previous literature indicates that relative advantage, complexity, and compatibility are more prominent factors influencing the diffusion of innovation (Kang et al., 2015; Agarwal & Prasad, 1998). We anticipate that for people to switch to using drove delivery service, the complexity of the technology itself (e.g., navigation, licensing, etc.) is hidden from immediate view from the consumer. Therefore, for the purposes of this study, we aim to look specifically at relative advantage and compatibility, as they are visible and more tangible to the e-commerce consumer.

Switching Behaviors

Switching intention refers to the decision of users to abandon current services and embrace new services (Bansal et al., 2005). Research has investigated a multitude of factors that have influenced people's switching intentions. For example, Sun (2013) proposed a new construct of "imitating others" in his longitudinal study of new technology adoption based on a herd behavior perspective.

Herd behavior has been witnessed in many consumers' behaviors in situations involving uncertainties and risks. For example, Keynes (1930, 1936) and other economists, including Minsky (1975) and Kindleberger & Aliber (2005), explained financial herd behaviors in the stock market as the outcome of the sociological and psychological forces in uncertain times. According to them, uncertainty encourages people to believe what others believe and do what others do. Studies involving technology also identify the same behavior patterns in using a new software application (Duan et al., 2009).

Imitating others means that "a person who is herding observes others and makes the same decisions or choices that others have made" (Sun, 2013). It differs from subjective norm - one of the most adopted constructs in IS research. Social norm (Fishbein and Ajzen, 1975; Davis et al., 1989; Venkatesh et al., 2003) refers to a person's perception of what others think they should or should not do. So, motivation-wise, imitating others meant avoiding costs or mistakes rather than being concerned about social impressions.

Gen Z

Generation generally refers to individuals born and raised at a similar time. Generation Z, or post-millennials, was born in the 1990s and raised in the 2000s (Pew Research Center, 2019). The specific social and economic developments associated with time will likely generate characteristics, attitudes, values, and capabilities unique to each generation (Berkup, 2014). Since the 1990s, personal computers and internet technologies have profoundly changed societies. Generation Z is born and brought up in a world filled with the internet, laptops, smartphones, Wi-Fi, digital media, and social networks (Bascha, 2011). New technologies are a part of the natural environment for Generation Z. They have grown to be the most technologically sophisticated generation and are called digital natives, .com generation, iGen, etc. (Levickaite, 2010).

Growing up with technologies, Generation Z has developed a high dependency on technology and is instant-minded (Singh & Dangmei, 2016). Speed addiction is part of Generation Z's most distinctive traits (Berkup, 2014). They want anything to happen quickly and instantly. Therefore, the faster speed offered by drone delivery is likely to be highly attractive and regarded by them. Second, although technology is part of Generation Z's identity, and they are tech-savvy, Generation Z is still in their teens or early 20s and reaching maturity. The excessive exposure and use of technologies have repercussions. For example, in a quest for social affiliation and virtual bonding, they can be less concerned about privacy and continuously share too much personal information on social platforms (PrakashYadav & Rai, 2017). They lack problem-solving skills and have not demonstrated the critical thinking to check a situation in context, analyze it, and decide (Joseph Coombs, 2013). Social influence becomes a significant predictor of their adoption of technologies such as e-Books (Srirahayu et al., 2021) and m-commerce (Meghisan-Toma et al., 2021) and their purchasing behaviors (Kahawandala et al., 2020). Generation Z likely demonstrates the same behavioral tendency in their switching intention to innovative technology-enabled drone delivery by imitating peers. Yet, there is little research about how these digital natives perceive and react to drone delivery.

3. HYPOTHESE DEVELOPMENT

Relative Advantage of Drone Delivery.

Many users are attracted to the relative advantages of delivery speed (Kornatowski et al., 2018). Rogers (1983) asserts that the critical driver for the diffusion of innovation is the relative advantages of the innovation. Previous research indicates that the more perceived advantages, the higher the likelihood of innovation adoption (Agarwal & Prasad, 1998; Kang et al., 2015). Studies and surveys suggest that innovative drone delivery provides a significant benefit - delivery speed (Yoo et al., 2018). Online shoppers perceive speedy delivery as the main advantage because drones fly over ground obstacles and in the optimal path and are not affected by road infrastructure or traffic congestion (Joerss et al., 2016). The packages can be delivered at the desired time because delivery time can be correctly predicted (Joerss et al., 2016). Accordingly, this study proposes that the perceived advantages of speedy delivery will lure online shoppers into switching from standard truck services to drone delivery.

H1: the relative advantage of delivery speed increases the switching intention from standard truck delivery to drone delivery.

Compatibility of Drone Delivery.

Tornatzky and Klein (1995b) find that innovations are more likely to be adopted when they are compatible with individuals' job responsibilities and value systems. Rogers (1983) defines compatibility as the degree to which an innovation meets the needs of potential customers with experience and existing values.

Extant research shows a positive association between compatibility and adopting new technologies such as Uber (Min et al., 2019). This research argues that if online shoppers perceive drone delivery to be compatible with their lifestyles and values, shopping preferences and habits, and new technologies acceptance attitudes, they are likelier to switch delivery services. Hence, we propose:

H2: the perceived compatibility increases the switching intention from standard truck delivery to drone delivery

Imitation of Drone Delivery
Sun's (2013) study suggests that imitation can help reduce post-adoption regret for making a choice; thus, potential adopters legitimate it as an effective strategy to choose a technology. Sun also points out that the two primary factors influencing "imitating others" are behavior observability and perceptions of uncertainty about the new technology. The previous discussion mentioned that observability is inapplicable in drone delivery due to unavailability. Thus, uncertainty is the main drive for online shoppers to imitate others in drone delivery. In other words, "imitating others" helps mitigate the uncertainty perception of innovations. Thus, we propose:

H3: imitating others positively affects switching intention from standard truck delivery to drone delivery.

Figure 1 illustrates the final research model for this research with hypotheses.

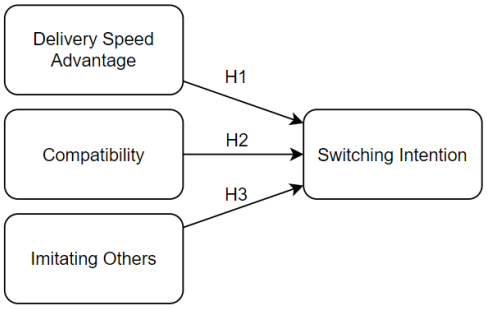


Figure 1. Research Model

Data Collection Procedure
This research studies the emerging character of drone delivery and how it affects people's perceptions about switching to such services. Our experimental study asked the respondents to watch a video about drone delivery first. A scenario describes a possible future situation, including the path of development leading to that situation (Kosow & Gabner, 2008). Scenarios are not intended to represent a complete description of the future but rather to highlight a possible future's central elements and draw attention to the key factors. Our scenario asks respondents to step into the situation where researchers want them to be and answer the question, "what would you do in this scenario" (Bishop et al., 2007).

4. METHODOLOGY

Table 1: Demographic Sample Statistics

Sample Size:	N = 83
Gender	*%*
Female	34.9
Male	65.1
Transportation Access	*%*
Yes	92.8
No	7.2
Commercial Drone Proximity	*%*
Yes	4.8
No	7.2
Amazon Prime Membership	*%*
Never had	7.2
Formerly, not currently	16.9
Current member	75.9
Current Neighborhood	*%*
Extremely close (tight)	32.5
Very close	36.1
Moderately close	27.7
Not close (distant)	3.6
Frequency of deliveries	*Avg*
Per month	1.8
Distance to the closest shopping center	*Avg*
Minutes	7.1

Structural equation modeling (SEM) with SmartPLS and a PLS algorithm is used to test and analyze the hypotheses of our reflective research model (Hair et al., 2017). Partial Least Squares SEM allows us to explore and estimate hypothesized complex predictive relationships between latent constructs (Hair et al., 2017). The advantages of no assumption on normal data distribution and the model convergence on a relatively small sample size fit our proposed exploratory theory (Hair et al., 2017).

We collected our survey data from 83

undergraduate students at a business school. This group best represents Generation Z, our target demographic in understanding perceptions of drone delivery. All responses were recorded on a 7-point strongly disagree (1) – strongly agree (7) Likert scale except for the switching intention construct, which has multi-item semantic-differential scales. Different scales within the same survey questionnaire help lower common method bias, as suggested by research (e.g., Podsakoff et al., 2003; Heppner et al., 2008). In addition, manipulation questions such as speeder trap and attention filter are used to eliminate common method bias further (Oppenheimer et al., 2009; Meade & Craig, 2012; Berinsky et al., 2014). In the beginning, a draft of the adapted items was reviewed and pretested within a group of students consisting of six graduate and four undergraduate students. The items' wording and organization were revised based on the feedback to ensure clarity in the drone delivery context. Next, an online pilot survey collected 50 responses. The proposed original research model with the relative advantage in speed, compatibility, complexity, trialability constructs of diffusion of innovation theory, and additional imitating others construct was tested while observability was not because it is not applicable in the context. The insignificant results suggest eliminating the constructs of complexity and trialability in the context of drone delivery. Instruments were also fine-tuned further based on the results from the data collected. At last, the survey questionnaire with 24 items, including questions to capture the demographics of respondents and usage patterns in shopping, was finalized and used to collect 83 effective responses. Table 1 shows the demographic snapshot of respondents of the main study.

Survey Instruments

The constructs in this study were measured using items adapted from previously validated studies (Appendix 1). We adopt three likelihood semantic items from Bansal et al. (2005) to gauge the switching intention to drone delivery. For example, respondents were asked to rate the chance of switching to drone delivery, such as "the likelihood that you would switch from truck delivery to drone delivery." There are four items to measure the attitude of respondents towards the advantages of delivery speed offered by drone delivery. These items were initially designed to test the emerging technology diffusion process (Moore & Benbasat, 1991). Extant research has applied them to various technology-enabled service contexts such as mobile banking (e.g., Al-Jabri & Sohail, 2012). Imitating others was adapted from Sun (2013). We removed the three reverse-coded items after the first round of the pilot study due to the inconsistent survey results. Research indicates that reverse-worded items failed to prevent response bias and contaminated data due to respondent inattention and confusion (Sonderen et al., 2013). The final constructs and their associated definitions are presented in Table 2.

Table 2. Construct Definitions

Construct	Definition
Delivery Speed Advantage (ADVS)	The improved relative advantage offered by drone delivery service over traditional truck delivery.
Compatibility (COMP)	The harmonious offering was given by drone delivery service with respect to traditional truck delivery.
Imitating Others (IMI)	The duplication of others' observed actions in the use of drone delivery service.
Switching Intentions (SWINT)	The committed resolve to use drone delivery service over traditional truck delivery (when available).

5. RESULTS AND DISCUSSION

Measurement Model

The measurement model estimates the accuracy of variables (measurement items), the relationships between the measured variables, and the latent constructs they represent. This involves assessing and evaluating items' loadings, construct's composite reliability, convergent and discriminant validity, and overall measurement model fit. Table 3 provides the final operationalized items loadings, and Table 4 the descriptive statistics of each construct.

Nunnally (1978) suggests that composite reliability should be 0.7 or higher for a construct to demonstrate adequate reliability. Convergent validity refers to the extent to which items for each construct are related and measures the same construct, evaluated by average variance extracted (AVE). A larger than 50% variance in each construct is suggested (Hair et al., 2009). Table 5.

Table 3. Outer Model Loadings

Item	ADVS	COMP	SWINT	IMI
ADVS1	0.841			
ADVS 2	0.905			
ADVS 3	0.888			
ADVS 4	0.808			
COMP1		0.728		
COMP2		0.781		
COMP3		0.806		
COMP4		0.824		
COMP5		0.717		
IMI 1			0.858	
IMI 2			0.916	
IMI 3			0.816	
IMI 4			0.825	
SWINT 1				0.964
SWINT 2				0.956
SWINT 3				0.938

In contrast, discriminant validity ensures that variables of each construct are not interrelated and only measure their associated constructs. It can be evaluated using a Fornell-Larcker criterion and a heterotrait-monotrait ratio of correlations (HTMT) in SmartPLS. The Fornell-Larcker values (square root of every AVE), reported in bolded font and the diagonal of the correlation matrix (Table 4), are larger than the corresponding off-diagonal correlations among any pair of latent constructs (Fornell & Larcker, 1981), indicating suitable discriminant validity. The HTMT is a new method outperforming classic approaches to discriminant validity assessment (Voorhees et al., 2016; Henseler et al., 2015). The values (Table 5) are smaller than 1, indicating good discriminant validity (Ab Hamid et al., 2017; Kline, 2011).

Table 4. Descriptive statistics for each construct

Construct	No. of Items	Mean	Std Dev.
Delivery Speed Advantage (ADVS)	4	5.12	1.13
Compatibility (COMP)	5	4.90	1.41
Imitating Others (IMI)	4	4.27	1.34
Switching Intentions (SWINT)	3	3.24	1.53

The overall standardized root mean square residual (SRMR) measures the model's residual discrepancies between observed and hypothesized correlations. Our SRMS (.090) is at par with the suggested cut-off values (Hu & Bentler, 1999; Byrne, 2016; Kline, 2011), demonstrating a good model fit (McDonald & Ho, 2002).

Nunnally (1978) suggests that composite reliability should be 0.7 or higher for a construct to demonstrate adequate reliability. Convergent validity refers to the extent to which items for each construct are related and measures the same construct, evaluated by average variance extracted (AVE). A larger than 50% variance in each construct is suggested (Hair et al., 2009). Table 5.

Table 5. Latent Variable Correlations, AVE, CR, and Chronbach's alpha (reliability)

Variable	1	2	3	4	AVE	CR	α
1. Delivery Speed Advantage (ADVS)	[0.861]				0.742	0.967	0.884
2. Compatibility (COMP)	0.541	[0.772]			0.742	0.920	0.884
3. Imitating Others (IMI)	0.422	0.510	[0.855]		0.731	0.915	0.883
4. Switching Intentions (SWINT)	0.483	0.565	0.488	[0.953]	0.597	0.881	0.830

Note:
Model Fit Statistics: SRMR = .090, χ^2 = 287.4;
AVE: average variance extracted, CR: composite reliability, α: Cronbach's alpha, N = 83
√AVE represented on diagonal in [ ]
Confirmed both discriminant and convergent validity using Fornell and Larcker (1981) method

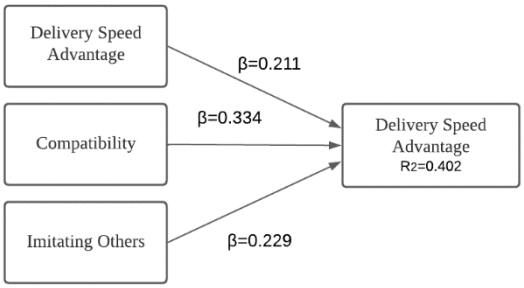


Figure 2. Model Results

Structural Model and Hypotheses Testing
Figure 2 and Table 6 summarize the model results, including the standardized path coefficients for each hypothesized relationship and associated ρ values. Our overall model's R^2, representing the variance explained in the exogenous construct Switching Intention is 0.402, a reasonable outcome measuring the model's predictive accuracy (Hair et al., 2011; Chin, 1998; Falk & Miller, 1992). Testing H1, the delivery speed advantage (relative advantage) resulted in a significant relationship to switching intention (β = 0.211, ρ = 0.049<0.05), showing support for this hypothesis. The direct relationship of compatibility is also found to be significant (β = 0.334, ρ = 0.002<0.01), supporting H2. The final relationship between imitation and switching intention is significant (β = 0.229, ρ = 0.012<0.05), supporting H3.

Table 6. Hypotheses Summary

Hypothesis	Supported
H1: ADVS -> SWINT	YES
H2: CPAT -> SWINT	YES
H3: IMI -> SWINT	YES

Demographic Variations in Structural Model
For a study examining the interdisciplinary phenomenon, introducing controlling variables to check the influence of other extraneous or confounding variables can help enhance the model's internal validity (Tucker & Roth, 2006). Retailer proximity and shopping frequency are crucial factors that affect customer loyalty (Gahinet & Cliquet, 2018). However, this study did not find significant effects of the two variables. P-value is 0.984 for shopping center distance and 0.461 for delivery frequency, with a 0.004 R^2 increase in the controlled model. Comparative analysis of drone delivery switching intention among demographics is strategically critical to understanding societal factors associated with attitudes and developing marketing campaigns. Hence, this research also runs multigroup comparisons across gender, AmazonPrime membership, and neighborhood styles. The group with less than ten respondences is excluded to ensure statistical power (Hair et al., 2017).

6. IMPLICATIONS, FUTURE RESEARCH, AND LIMITATIONS

Theoretical implications
Our study contributes to innovation diffusion and imitation as theoretical lenses. First, this study builds upon the diffusion of innovation theory that can help understand the reasons causing new technology adoption (E. Rogers, 1995; E. M. Rogers, 2010). Our study shows that speed and lifestyle compatibility are two variables critical to the increased intention of users to switch to drone delivery services. This finding confirms the importance of relative advantages and compatibility as two crucial attributes of innovations in drone delivery services. Thus, our finding increases the applicability and generalizability of the innovation diffusion theory to drone delivery services.

Second, unlike most innovation diffusion studies using adoption intention as the dependent variable (Salahshour Rad, Nilashi, & Mohamed Dahlan, 2018), this study adopts switching intention as the dependent variable. The dependent variable is the switching intention rather than the adoption intention, commonly used in the innovation diffusion theory. Using the unconventional dependent variable is another theoretical contribution to the current innovation diffusion literature.

Third, this study adopts the imitation theory (Kinnunen, 1996; Tarde, 2013) and incorporates the imitation variable into the research model. Our finding shows that users' decision to switch to drone delivery services depends on the suggestion-imitation process. The process is one of the primary reasons for the proliferation of social networks (Gibbs, 2008).

Furthermore, this finding corroborates previous studies on the influence of the imitation process on social movements, including adopting new technology and communication channels (Lee, Trimi, & Kim, 2013). The addition of imitation as the antecedent for the switching intention advances our understanding of drone delivery services. Increasing users' intention to switch to drone delivery services needs to improve technology capability, computability, and

imitation process.

Practical implications

Gen Z-ers are particularly fascinated by the significant technological improvement in delivery speed over traditional trucking services. Suppose e-commerce providers, such as Amazon, can deliver goods in a much shorter time frame (30 minutes for Prime Air services) than the current delivery speed (one to two days). In that case, Gen Z-ers are motivated to switch to drone delivery services. Therefore, e-commerce providers should focus on improving their own logistics systems or partner with logistics service providers (e.g., UPS and FedEx) to continuously improve the velocity of drone delivery services.

Drone delivery fits the mobility and technology-saturated lifestyle of Gen Z-ers. Generation Z or Gen Z are people born between the mid-late 1990s and the early 2010s. Gen Z-ers are digital natives accustomed to using social media, smartphones, apps, and other emerging technologies in their everyday lives (Kurzu, 2017). The mobility and technology-saturated lifestyle enable Gen Z-ers to learn new technologies proactively. Drone delivery services are another new technology compatible with Gen Z-ers' lifestyles. E-commerce vendors can focus on recruiting Gen Z-ers who are innovators or early adopters to experiment with drone delivery services. If these pioneers love drone delivery services, they will quickly inform and influence others to embrace them.

Observability or visibility is indispensable for the diffusion of innovation (Magsamen-Conrad & Dillon, 2020). Users are more likely to adopt or switch to an unproven technology if they see more people using it. Our study shows that the imitation process is the result of observability. Users who see others using drone delivery services are more likely to embrace it while moving away from the current delivery method. The leading short video platforms, such as Tiktok, have leveraged the mimesis logic and design to encourage viewers to imitate and replicate popular videos to alter modes of sociality (Zulli & Zulli, 2020). E-commerce vendors may want to leverage the social community or social influencers to improve the visibility of drone delivery services. The increased visibility can ultimately result in the increased intention of users to switch to drone delivery services.

Gen Z-ers grew up understanding new technologies and their technical benefits. This new generation of users is more receptive to unfamiliar technologies, such as drone delivery services. Our study suggests that Gen Z-ers are comfortable switching to drone delivery services if they can fit (compatibility) in their mobility lifestyle and deliver on their promise. E-commerce vendors and logistics service providers may be able to take advantage of a ripe market and attempt to overcome any technical glitches or barriers to innovating better drone delivery services.

Limitations and future research directions

The study has several limitations even after conducting a rigorous design and control of survey instruments and data analysis. First, this study surveyed only American subjects who had never used drone delivery services. Although respondents did spend time watching two short videos and learning about the benefits and risks of drone delivery services, they lacked real-life experiences of using them in their daily lives. Therefore, the findings based on these limitations may not generalize to other non-American subjects and users who have already had experiences using drone delivery services. Future research can survey users from different countries (e.g., Australia, China, and the United Kingdom) with varying drone delivery experiences.

Second, all subjects surveyed in this study are college students from a regional American university. These students are the best candidates for Gen Z-ers because they are educated and affluent with using different technologies to maintain their mobility and technology-saturated lifestyle. Surveying these subjects can provide a realistic observation of drone delivery adoption behaviors. Although Gen Z-ers are an excellent segment for drone delivery services, the findings of this study cannot be generalizable to other customer segments who are non-Gen Z-ers. Future research can also collect data from other populations to provide more insights into diverse drone delivery switching behaviors.

Third, this study closely examines the influence of two critical predictors for the rate of innovation adoption: relative advantage and compatibility (Min, 2019). However, this study did not examine the other critical elements of any innovation adoption: complexity, observability, and trialability. This study purposely eliminated these elements because respondents did not have prior knowledge or experience using drone delivery services. Future research can include these elements in the survey design by expanding this study with subjects with the experience of using drone delivery services.

Fourth, this study selects two short videos to ensure that subjects have exposure to the same information regarding drone delivery services. The questions used in the survey may not be able to fit perfectly with the content of the selected videos. As a result, the content and quality of these videos may have influenced the survey results. However, the current research design could not filter out the uncontrollable bias. Future research can continuously look for videos more compatible with survey questions.

Fifth, this study developed an integrative research model by combining innovation diffusion and imitation theories. The factors derived from these theories provide limited observations about the switching behaviors of Gen Z-ers to drone delivery services. Although the integrative research model can explain 40% of the switching intention variability, other factors can further improve the predictive power. Future research can consider adopting an interdisciplinary perspective to better understand the switching behaviors of drone delivery services.

Last, our findings are about users' perceptions of switching to drone delivery services. Future research can use the qualitative research methodology for interviewing users to gain first-hand experiences using drone delivery services in everyday life (Kaufmann, Peil, & Bork-Hüffer, 2021). The qualitative approach can help cross-examine the findings of this study and provide richer observations of drone delivery switching behaviors.

7. REFERENCES

Al-Jabri, I. M., & Sohail, M. S. (2012). Mobile banking adoption: Application of diffusion of innovation theory. *Journal of electronic commerce researc*h, *13*(4), 379-391.

Alwateer, M., & Loke, S. W. (2020). Emerging drone services: *Challenges and societal issues. IEEE Technology and Society Magazine*, *39*(3), 47-51.

Barth, S., & De Jong, M. D. (2017). The privacy paradox–Investigating discrepancies between expressed privacy concerns and actual online behavior–A systematic literature review. *Telematics and Informatics*, *34*(7), 1038-1058.

Chatterjee, S., Chaudhuri, R., Vrontis, D., & Hussain, Z. (2021). Usage of smartphone for financial transactions: from the consumer privacy perspective. *Journal of Consumer Marketing*.

Damanpour, F. (1991). Organizational innovation: a meta-analysis of effects of determinants and moderators, *Academy of Management Journal*, *34*(3), 550-90.

Gahinet, M.-C., & Cliquet, G. (2018). Proximity and time in convenience store patronage: Kaïros more than chronos. *Journal of Retailing and Consumer Services*, 43, 1-9.

Gibbs, A. (2008). Panic! Affect contagion, mimesis and suggestion in the social field. *Cultural Studies Review*, *14*(2), 130-145.

Hassandoust, F., Akhlaghpour, S., & Johnston, A. C. (2021). Individuals' privacy concerns and adoption of contact tracing mobile applications in a pandemic: A situational privacy calculus perspective. *Journal of the American Medical Informatics Association*, *28*(3), 463-471.

Hsieh, P. J., & Lin, W. S. (2020). Understanding the performance impact of the epidemic prevention cloud: an integrative model of the task-technology fit and status quo bias. *Behaviour & Information Technology, 39*(8), 899-916.

Ifinedo, P. (2012). Understanding information systems security policy compliance: An integration of the theory of planned behavior and the protection motivation theory. *Computers & Security*, *31*(1), 83-95.

Kaufmann, K., Peil, C., & Bork-Hüffer, T. (2021). Producing In Situ Data from a Distance with Mobile Instant Messaging Interviews (MIMIs): Examples From the COVID-19 Pandemic. *International Journal of Qualitative Methods*, 20, 16094069211029697.

Kinnunen, J. (1996). Gabriel Tarde as a founding father of innovation diffusion research. *Acta sociologica*, *39*(4), 431-442.

Kornatowski, P. M., Bhaskaran, A., Heitz, G. M., Mintchev, S., & Floreano, D. (2018). Last-centimeter personal drone delivery: Field deployment and user interaction. *IEEE Robotics and Automation Letters*, *3*(4), 3813-3820.

Kurzu, R. (2017). *Generation Z: The Lasting Influence of the Digital Native on Marketing*.

Lee H-J, Yang K. (2013). Interpersonal service

quality, self-service technology (SST) service quality, and retail patronage. *Journal Retail Consumer Services*, *20*(1), 51–57

Lee, H. L., Chen, Y., Gillai, B., Rammohan, S. (2016). Technological Disruption and Innovation in Last-Mile Delivery. Retrieved from Stanford Graduate School of Business: https://www.gsb.stanford.edu/faculty-research/publications/technological-disruption-innovation-last-mile-delivery.

Lee, S. G., Trimi, S., & Kim, C. (2013). Innovation and imitation effects' dynamics in technology adoption. *Industrial Management & Data Systems*.

Li, H., Wu, J., Gao, Y., & Shi, Y. (2016). Examining individuals' adoption of healthcare wearable devices: An empirical study from privacy calculus perspective. *International Journal* of *Medical Informatics*, *88*, 8-17.

Magsamen-Conrad, K., & Dillon, J. M. (2020). Mobile technology adoption across the lifespan: A mixed methods investigation to clarify adoption stages, and the influence of diffusion attributes. *Computers in Human Behavior*, *112*, 106456.

Milne, G. R., Rohm, A. J., & Bahl, S. (2004). Consumers' protection of online privacy and identity. *Journal of Consumer Affairs*, *38*(2), 217-232.

Min, H. (2019). Blockchain technology for enhancing supply chain resilience. *Business Horizons*, *62*(1), 35-45.

Moore, G.C., Benbasat, I. (1991). Development of an instrument to measure the perceptions of adopting an information technology innovation. *Information Systems Research*, *2*(3), 192–222.

Rogers, E. M. (1995). Diffusion of innovations–5th edition Free Press. New York.

Rogers, E. M. (2010). Diffusion of Innovations: Simon and Schuster.

Rogers, R. W. (1975). A protection motivation theory of fear appeals and attitude change. *Journal of Psychology*, *91*(1), 93–114.

Salahshour Rad, M., Nilashi, M., & Mohamed Dahlan, H. (2018). Information technology adoption: a review of the literature and classification. *Universal Access in the Information Society*, *17*(2), 361-390.

Schlee, R. P., Eveland, V. B., & Harich, K. R. (2020). From Millennials to Gen Z: Changes in student attitudes about group projects. *Journal of Education for Business*, *95*(3), 139-147.

Sıcakyüz, Ç., & Yüregir, O. H. (2020). Exploring resistance factors on the usage of hospital information systems from the perspective of the Markus's Model and the Technology Acceptance Model. *Journal of Entrepreneurship, Management and Innovation*, *16*(2), 93-131.

Sinkula, J.M., Baker, W.E. and Noordewier, T. (1997), A framework for market-based organizational learning: linking values, knowledge and behavior. *Journal of the Academy of Marketing Science*, *25*(4), 305-18.

Soffronoff, J., Piscioneri, P., Weaver, A. (2016). Public Perception of Drone Delivery in the United States (RARC-WP-17-001). Retrieved from U.S. Postal Service Office of Inspector General https://www.uspsoig.gov/document/public-perception-drone-delivery-united-states.

Spachos, P., & Gregori, S. (2019). Integration of wireless sensor networks and smart UAVs for precision viticulture. *IEEE Internet Computing*, *23*(3), 8-16.

Statista. (2022). Share of online consumers in the United States who are Amazon Prime members in 2019, by generation. Retrieved from https://www.statista.com/statistics/609991/amazon-prime-reach-usa-generation/

Sun, H. (2013). A longitudinal study of herd behavior in the adoption and continued use of technology. *MIS Quarterly*, 1013-1041.

Sun, H., and Fang, Y. (2010). Toward a Model of Mindful Acceptance of Technology. In *Proceedings of the 31st International Conference on Information Systems*, St. Louis, MO, December 12-15.

Tan, M., Teo, T.S., (2000). Factors influencing the adoption of Internet banking. *Journal of the Association for Information Systems*, *1*(5), 1–42.

Tarde, G. (2013). *The laws of imitation*, Read

Books Ltd.

Thong, J. Y. L. (1999). An Integrated Model of Information Systems Adoption in Small Businesses. *Journal of Management Information Systems*, *15*(4), 187–214. https://doi.org/10.1080/07421222.1999.11518227

Tucker, J. A., & Roth, D. L. (2006). Extending the evidence hierarchy to enhance evidence-based practice for substance use disorders. *Addiction*, *101*(7), 918-932.

Vance, A., Siponen, M., & Pahnila, S. (2012). Motivating IS security compliance: insights from habit and protection motivation theory. *Information & Management*, *49*(3-4), 190-198.

Yaprak, Ü., Kılıç, F., & Okumuş, A. (2021). Is the Covid-19 pandemic strong enough to change the online order delivery methods? Changes in the relationship between attitude and behavior towards order delivery by drone. *Technological Forecasting and Social Change*, 169, 120829.

Yoo, W., Yu, E., & Jung, J. (2018). Drone delivery: Factors affecting the public's attitude and intention to adopt. *Telematics and Informatics, 35*(6), 1687-1700.

Zhu, X., Pasch, T. J., & Bergstrom, A. (2020). Understanding the structure of risk belief systems concerning drone delivery: A network analysis. *Technology in Society*, *62*, 101262.

Zulli, D., & Zulli, D. J. (2020). Extending the Internet meme: Conceptualizing technological mimesis and imitation publics on the TikTok platform. *New Media & Society*, 1461444820983603.

APPENDIX 1. MEASUREMENT ITEMS

A1. Construct Measurement Items

Construct	Item	Reference
Relative Advantage (7-point Likert)	Drone delivery is a speedy way to deliver orders	Al-Jabri et al., 2012; Moore and Benbasat, 1991
	Drone delivery allows me to receive orders quickly	
	Drone delivery is useful in shortening the order delivery time	
	Drone delivery gives greater control over the speed of my delivery	
Compatibility (7-point Likert)	Drone delivery fits well with the way I like to manage my online order delivery	Al-Jabri et al., 2012; Moore and Benbasat, 1991
	I like to try new technologies	
	Drone delivery is compatible with my lifestyle	
	Using drone delivery fits into my online shopping style.	
Imitating Others (7-point Likert)	It seems that drone delivery is the future dominant shipping service, therefore I would like to use it as well	Sun, 2005
	I would follow others in accepting drone delivery	
	I would choose to accept drone delivery because many other people are using it	
	If I know that a lot of people have already accepted drone delivery. I might choose drone delivery.	
Switching Intention	Rate the likelihood that you would switch from truck delivery to drone delivery (Likely – Unlikely)	Bansai et al., 2005
	Rate the probability that you would switch from truck delivery to drone delivery. (Probable – Not Probable)	
	Rate the chance that you would switch from truck delivery to drone delivery (Certain – No Chance)	

The Effect of Mental Illness on Compensation for IT Developers

Alan Peslak
arp14@psu.edu
Department of Information Sciences and Technology
Penn State University
Dunmore, PA 18512

Wendy Ceccucci
wendy.ceccucci@qu.edu

Kiku Jones
kiku.jones@qu.edu

Department of Computer Information Systems
Quinnipiac University
Hamden, CT 06518

Lori N. K. Leonard
lori-leonard@utulsa.edu
Department of Accounting and Computer Information Systems
The University of Tulsa
Tulsa, OK 74104

Abstract

Nearly 20% of US adults suffer from some form of mental illness. The indirect effects of the COVID-19 virus have increased this number significantly. The effect of mental illness on ability to work effectively and efficiently has been studied extensively and the consensus is that mental illness has a stigma associated with it that reduces employment opportunities for those so afflicted. Our study reviews this assumption by analyzing compensation for information technology developers who self-identify as having one of many mental illnesses. A large sample set from Stack Overflow was used to compare compensation levels based on this self-identification of one or more mental illnesses to determine if there was any significant impact. Our results found that overall mental illness does reduce compensation levels for information technology developers but less than many other variables. Other demographic factors, including a lower level of education, female gender, and younger age, are more significant factors for lower developer compensation. There is also a small effect based on ethnicity.

Keywords: Mental Illness, Compensation, Gender, Education

1. INTRODUCTION

The COVID-19 pandemic has caused a rise in the number of mental illness cases worldwide (Xie, Xu, & Al-Aly, 2022). Mental illness encompasses a wide range of conditions including depression, anxiety disorders, schizophrenia, eating disorders and addictive behaviors. It can affect a person's mood, thinking and behavior (MayoClinic, 2022a). Mental health "includes our emotional, psychological, and social well-being. It affects how we think, feel, and act. It also helps determine how we handle stress, relate to others, and make healthy choices" (Center for Disease Control, 2022). While the terms mental illness and mental health are commonly used interchangeably, they are not the same. A person may have a diagnosed mental illness, but at times be in good mental health. On the other hand, a person may be experiencing poor mental health without a diagnosed mental illness.

In 2019, just prior to the COVID-19 pandemic, 19.86% of U.S. adults experienced some form of mental illness (Mental Health America, 2022). SingleCare (2022) conducted a national survey on mental health and coronavirus and found that 59% of people in the US stated that their mental health was impacted by the COVID-19 pandemic. In addition, according to a study by the World Health Organization (2022), the COVID-19 pandemic has resulted in a 25% increase in the prevalence of anxiety and depression worldwide.

Prior research has shown that social disadvantage, especially lack of material possessions, lower income, and financial difficulties, are often associated with mental illnesses (Reading, 2000, Lewis, 1998, and Welch, 1998). In addition, women have been found to be diagnosed with a mental illness more often than men (Mayo Clinic, 2022b). During the COVID-19 pandemic, one in six women were found to have symptoms of post-traumatic stress – a rate like that of other significant disasters (Lindau, Makelarski, Boyd, Doyle, Haider, Kumar, Lee, Pinkerton, Tobin, Vu, Wroblewski, & Lengyel, 2021).

During the pandemic, many companies were required to conduct business virtually and could not operate at full capacity, which necessitated reducing their workforce (US Bureau of Labor Statistics, 2022). While the technology industry was not immune to this issue, it was in a better position overall (Hylton, Ice, & Krutsch, 2022). The pandemic created additional responsibilities and added layers of complexity to the IT professional's workload. This added stress and pressure to many IT professionals in the field (Thompson, 2022). Prior to the pandemic, fifty-one percent of IT professionals indicated that they were already diagnosed with a mental illness (OSMI, 2016). As new positions and roles are created to meet the demands post-pandemic, it will be important to understand how mental illness impacts the IT industry.

This paper looks to answer the following research questions:

- What impact does mental illness have on compensation (i.e., income) in the IT industry?
- What is the impact of gender in the relationship between mental illness and compensation?
- What is the impact of education in the relationship between mental illness and compensation?
- Are there any variables that have more of an impact on compensation than mental illness?

The next section will provide a literature review om the effects of mental illness on, compensation, education, and gender. This is followed by the methodology, results, and conclusions.

2. LITERATURE REVIEW

Mental Illness & Compensation

A study by Kose (2020) found a causal and positive relationship between household income level and the mental health of individuals in Turkey. They also reported that Turkish females had lower mental health than Turkish males.

Gresenz, Sturm, and Tang (2001) studied the relationship between income inequality and mental disorder. They found that mental health worsens, and the probability of depression and anxiety increases as the income of the family decreases. Wildman (2003) found that one's financial status is a determinant of mental health, and one's income creates inequalities in health. Moreover, Strohschein (2005) studied school age children between the ages of four and fourteen. She found that low household income results in greater depression, with improvements or increases in household income resulting in reduced mental health issues in children.

In particular, the COVID-19 pandemic lockdown was found to affect mental health more in low-income households (Pieh, Budimir, and Probst, 2020). In Pieh, Budimir, and Pobst's study they looked at residents of Austria during the first four weeks of lockdown. They determined that stress levels were higher for low-income households, adults under 35 years of age, and women. Pieh et al. (2021) also studied residents in the United Kingdom. They found that individuals from low-income households, younger than 35 years of age, and women experienced greater mental health issue. Li et al. (2020) also found similar results. They found that people who had larger income losses were at higher risk for mental health issues and required more psychological care. Historically, socially disadvantaged groups, such as those with low income, have shown more psychiatric illnesses or conditions than socially advantaged groups (Purtle, 2020).

However, another study by Araya et al. (2003) looked at adults living in Chile and found there was a strong, inverse, and independent association between education and common mental illnesses. But income was not associated with the prevalence of common mental illnesses, after adjusting for other socioeconomic variables.

Mental Illness & Gender

Prior research studies have found that women tend to experience higher instances of mental illness (Chochrane, 1981, Hankin, 2001, Macintyre 1996). The Substance Abuse and Mental Health Services Administration conducted the 2020 National Survey on Drug Use and Health and found that mental illness of any kind was higher among females (25.8%) than males (15.8%) (National Institute of Mental Health, 2022).

Seedat et al. (2009) found that gender differences in mental illness are consistently observed in various countries from different regions of the world, but they found that gender differences decline when men and women have more equal roles in the society.

Existing mental illness was appeared to be exacerbated during the pandemic. Many individuals were hospitalized for long periods of time without the ability to see their families. Prior to the pandemic, Paulo da Silva Ramos, et al. (2022) surveyed hospitalized patients using the Hospital Anxiety and Depression Scale and the Beck's Anxiety Inventory. Results showed that women had a higher level of anxiety than men. However, there was no difference in the level of depression. As previously stated, women were found to be affected more than men by the pandemic in Austria, with higher stress levels recorded for women during the first four weeks of lockdown (Pieh, Budimir, and Probst, 2020). This was also found to be true in the United Kingdom (Pieh et al., 2021).

Prowse et al. (2021) examined the COVID-19 pandemic's effect on the stress and mental health of university students in Canada. They found female students to more negatively affected than male students in terms of academics, stress, mental health, and isolation.

Due to increasing cases of Covid-19 cases many states called for lockdowns. Adams-Prassl, et al. (2022) studied the effects of the lockdown measures on mental health. They determined that the stay-at-home orders led to an overall decrease in mental health. They found a 61% increase in the gender gap in mental health with the negative impact of the lockdown impacting women more. The authors also found a positive association with income and a university degree with mental health.

Mental Illness & Education

Previous research has shown a definitive relationship between educational outcomes and mental illness. The research suggests two main reasons for this relationship, social causation, and social selection. Social causation research suggests that the relationship is a causal one and education affects mental health (Kessler et al. 1995; Lantz et al. 2005; Mirowsky and Ross 2003; Ritsher et al. 2001; Schieman and Plickert, 2008). Whereas the research on selection suggests that preexisting mental health conditions inhibit an individuals' ability to obtain a high level of education.

The causal relationship subscribes those higher levels of education enhance people's skills, afford important structural advantages, and empower better coping mechanisms, which results in better mental health.

The social selection theory prescribes that the link between education and mental illness may stem from preexisting conditions. Those experiencing these mental health problems are more likely to experience school difficulties, including more absenteeism, higher rates of suspension and expulsion, lower grades and test scores, and greater high school dropout propensity (Bernstein, 1997, Diperna, 2002, Gutman, 2003, & Reid, 2004). Research suggests that functional impairments and/or the stigma and social exclusion that go along with mental health illness

are often the cause (McLeod, Uemura, and Rohrman 2012; Needham, Crosnoe, and Muller 2004).

Alternatively, a British study, by Lewis et al. (1993) revealed an interaction between social class and sex, and no independent association with educational achievement.

3. METHODOLOGY

To study the effects of mental illness on software developers' data from the 2021 Stack Overflow survey was used. Stack Overflow's annual Developer Survey is the largest and most comprehensive survey of people who code around the world. Each year, their survey questions cover a wide range of areas, from developers' favorite technologies to their job preferences. According to their website:

> "For almost a decade, Stack Overflow's annual Developer Survey held the honor of being the largest survey of people who code around the world. This year (2020), rather than aiming to be the biggest, we set out to make our survey more representative of the diversity of programmers worldwide. That said, the survey is still big. This year's survey was taken by nearly 65,000 people." (Stack Overflow, 2020)

The use of Stack Overflow is well established as a source for peer-reviewed journals including Barua, Thomas, and Hassan (2014), Asaduzzaman, Mashiyat, Roy, and Schneider (2013), and Treude and Robillard (2016). The Stack Overflow dataset consists of dozens of demographics, descriptive, and opinion questions about the state of programming today. The results were analyzed using IBM SPSS 27. It is important to note that the responses regarding mental health were self-reported and could also include responses that were self-diagnosed.

This survey was used as the starting point for our analysis. Over 88,420 survey responses were received, but the survey included data on respondents who were just hobbyists or non-professional users. In addition, this was an international survey and to eliminate international variations in salary this research just focused on US only developers. Finally, there were anomalies in compensation with unrealistically large compensation levels. Those reporting yearly salaries over $500,000 were excluded. To summarize, the initial data was filtered to only include the surveys from individuals who lived in the United States, identified as a developer by profession, and whose converted compensation was less than $500,000. All responses that were missing data or the responses were rather not say were removed from the dataset. After filtering, the dataset was still quite large with just over 7,500 responses.

SPSS 27 was used to perform a variety of statistical analyses to determine the impact of mental health and information technology compensation. The working hypothesis was that if mental health does affect employment opportunities, would this be manifested in compensation levels. The goal was also to compare the magnitude of the effect of mental health issues on compensation as compared to other demographic variables.

4. RESULTS

The survey asked the following question to ascertain the respondents' mental health:

Which of the following describe you, if any? Please check all that apply.

- I have a concentration and/or memory disorder (e.g., ADHD)
- I have a mood or emotional disorder (e.g., depression, bipolar disorder)
- I have an anxiety disorder
- I have autism / an autism spectrum disorder (e.g., Asperger's)
- None of the above
- In your own words

The results of this survey question are shown in Appendix A and Figure 1. Appendix A shows the count and average salary of all the different combinations of responses. As respondents can check more than one disorder there are many different combinations. Figure 1 on the other hand, shows the respondents' results for each of the response options. Given that respondents can check more than one disorder, the sum of the columns is larger than the total number of respondents. Approximately, 67% of the respondents indicated that they had no mental disorders, and 33% indicated that they had some type of mental disorder.

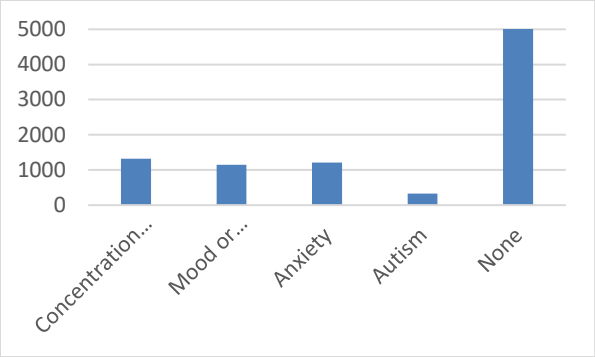


Figure 1: Count of Mental Disorder

Figure 2 shows the average salary of the respondents based on the mental disorder. The average salary of those indicating no mental disorder had the highest salary, just less than $140,000 while those who had an anxiety disorder had the lowest average salary of approximately $125,000.

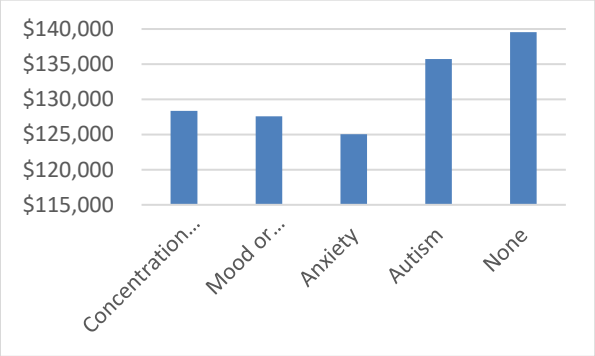


Figure 2: Average Salary by Mental Disorder

To further analyze if there was an overall difference between compensation levels for those who self-identified as having any mental illness versus those who did not, a simple binary variable was created. If the respondents indicated any type of mental disorder, the mental illness variable was a 1, otherwise it was set to zero. The results are shown in Tables 1 and 2. Overall, nearly 36% of the scrubbed dataset reported one or more mental illness. This is much higher than the estimated general population of 20%. As shown in table 2, there is more than a $10,000 difference between the compensation for those who reported a mental illness compared to those who did not. An ANOVA test was performed. This difference was significant a p< .001.

	Frequency	Percentage
No Mental Illness	5,011	66.5%
Mental Illness	2,520	33.5%
Total	7,531	

Table 1: Number of Respondents

	Mean	Std. Dev.
No Mental Illness	$139,547.99	$70,450
Mental Illness	$128,899.82	$65,125
Overall	$135,984.94	$68,893

Table 2: Average Salary by Mental Illness

To further investigate the factors that might affect the disparity in salary, each of the four important variables were added, that in past research have shown to affect job compensation: ethnicity, gender, age, and education level were separately analyzed.

Ethnicity was the first variable analyzed. Due to the more than 25 combinations of race in the dataset, the dataset was recoded to white and non-white so that a reasonable sample size was obtained. From the results in Table 3, the ethnicity factor used resulted in a small, but significant result for compensation. An ANOVA test was performed. The p value was less than 0.095, below the .10 threshold used in many social science studies. There was a $2000 increase in non-white versus white developers.

Ethnicity	Mean	N	Std Dev.
White	$135,276.90	5,863	$68,303
Non-White	$138,473.69	1,668	$70,895
Overall	$135,984.94	7,531	$68,894

Table 3: Average Salary by Ethnicity

Gender was the next variable that was independently analyzed. Those that self-reported and identified as other than solely male, or female were categorized into the other group. The results of the analysis are shown in tables 4 and 5. Women have significantly lower compensation. Their compensation was nearly $24,000 less than men and over $13,000 less than other gender. Women and Other also report much higher percentages of mental health issues (54% and 71%) versus men at 31%.

	Mean	N	Std Dev.
Male	$137,793.19	6,867	69,717
Female	$126,772.36	464	54,383
Other	$113,194.46	200	60,043

$p < .001$

Table 4: Average Salary by Gender

	No Mental Illness	Mental Illness	Total
Male	4,742	2,125	6,867
Female	212	252	464
Other	57	143	200

Table 5: Count of Gender & Mental Illness

The third independently reviewed variable was education level. The results are shown in Tables 6 and 7. In general, the higher the education level, the higher the compensation. An ANOVA test found this significant at $p < .001$. The results in Table 7 show that the level of education was affected by the having a mental disorder. An ANOVA test was performed, and it was determined that those having a higher education level had a lower incidence of mental illness. This was statistically significant at $p < .001$

Education	Mean	N	Std Dev.
Primary	$137,895.08	26	$81,273
High school	125,165.05	146	82,566
Some college	129,091.27	944	66,979
Associate	109,942.66	315	50,213
Bachelors	133,780.40	4472	67,586
Masters	151,218.19	1372	71,064
Doctoral /Advanced	156,296.00	256	76,581
Overall	135,984	7,531	68,893

Table 6: Average Salary by Education Level (Scaled)

Education	No Mental Illness	Mental Illness	Total
Primary	15 (58%)	11 (42%)	26
High school	80 (55%)	66 (45%)	146
Some college	507 (54%)	437 (46%)	944
Associate	181 (57%)	134 (43%)	315
Bachelors	3,009 (67%)	1,463 (33%)	4,472
Masters	1,041 (76%)	331 (24%)	1,372
Doctoral /Advanced	178 (70%)	78 (30%)	256
Overall	5,011 (67%)	2,520 (33%)	7,531

Table 7: Mental Illness & Education Level

The last factor analyzed was age. Age also had a significant impact on compensation level as shown in Tables 8 and 9. In general, older developers, at least up to age 54, had a higher compensation level than their younger counterparts. Also, younger developers from the appear to have higher levels of mental health issues as shown in Table 9. An ANOVA test was performed. Age was determined to be a significant variable affecting compensation at $p < .001$.

Age (years)	Mean	N	Std. Dev
18-24	$89,901.14	839	48382
25-34	130,281.01	3358	67099
35-44	153,857.37	2039	70,828
45-54	154,771.65	836	66,698
55-64	150,384.71	392	63,263
65 or older	136,370.15	67	60,518
Total	135,984.94	7,531	68,893

Table 8: Average Salary by Age

Age (years)	No Mental Illness	Mental Illness	Total
18-24	547 (65%)	292 (35%)	839
25-34	2,161 (64%)	1,197 (36%)	3,358
35-44	1,355 (66%)	684 (34%)	2,039
45-54	593 (71%)	243 (29%)	836
55-64	300 (77%)	92 (23%)	392
65 or older	55 (82%)	12 (18%)	67

Table 9: Age and Mental Health

With all these variables, different combinations will produce different results but to see an actual impact in a group, a major subset of the overall dataset was analyzed. The group consisted of only non-Women, Bachelor's degree, age 25-34, and white ethnicity. It was found that there was a $7000 difference in compensation for those without mental illness versus those with mental illness (Table 10). An ANOVA test was performed. This difference was significant at $p < .046$.

Mental Illness	Mean	N	Std. Dev.
No Mental Illness	$131,909.38	1,014	66,477
Mental Illness	124,982.51	543	62,625
Total	129,493.66	1,557	65,223

Table 10: Salary Differential of non-Women, Bachelor's degree, age 25-34, and white ethnicity with Mental Health

The final step was to utilize all the significant variables in a multiple regression analysis to determine the possible significance of each variable as well as determine any possible collinearity effects. The results of this analysis are shown in Appendix B.

Before discussing the specific regression coefficients and significance levels the collinearity analysis is first addressed. "Collinearity refers to the non-independence of predictor variables, usually in a regression-type analysis. It is a common feature of any descriptive ecological data set and can be a problem for parameter estimation because it inflates the variance of regression parameters and hence potentially leads to the wrong identification of relevant predictors in a statistical model." (Dorman, et al., 2013, p. 27) In other words, collinearity is bad and can result in improper conclusions because the variables may be non-independent and thus inaccurately predict the dependent variable.

There are multiple methods to determine collinearity and the two main methods are included in the SPSS output tables. The three key indices are Tolerance, VIF (Variance Inflation Factor) and Condition Index. The negative thresholds for these factors are less than .1 for tolerance, greater than 10 for VIF and greater than 30 for Condition index (Dorman, et al., 2013). The results show for all the variables the VIF's are below 1.1, all tolerances greater than .9 and no dimension condition indices above 16. Therefore, we can conclude that there are no collinearity issues with the variables, and all are independent of each other.

Now that the collinearity was checked, the next step was to analyze the results. It is worth noting that all variables were scaled or were dichotomous to allow for regression analysis. For gender, two dummy variables were added to determine the effect of women versus men and other versus men.

All variables included in the regression analysis were statistically significant at $p < .002$ except for other gender. Other gender did not have an impact on compensation with a p value of .20.

Evaluating all the variables, the statistical results show:

- Age is a significant factor in determining developer compensation. The older you are up to a point, the higher your compensation.
- Gender is a significant variable but only for women, not other. Women have lower compensation than men.
- Education level is a significant factor in determining developer compensation. The more education, the higher the compensation.
- Mental health is a significant factor by itself in determining developer compensation.
- Finally, ethnicity is also significant factor in determining developer compensation.

But these variables do not equally affect compensation. A review of the standardized coefficients reveals that the most important variable is age at .22. The next most important variable is Education level at .095. The third most important variable is gender but only for women at -.075. Ethnicity also has a slightly higher impact at .036, higher than mental health which is the lowest at -.035.

This data also supports the need for transparency in salaries to help ensure pay equity across all dimensions including ethnicities. Practitioners should work towards correcting their salary disparities between ethnicities.

5. CONCLUSIONS AND LIMTIATIONS

The data suggests that overall mental health issues are a significant factor in determining IT developer compensation, but not as high as other variables. Rather age and education level appear to be more important factors in contributing towards lower compensation. In addition, female gender, and ethnicity also play stronger roles. So overall, independently, mental health issues do reduce compensation, but have less effect than many other variables.

It should be noted that the data was from respondents who self-diagnosed their mental health. Additional research should be performed with medical data to further support these results.

Additionally, this study looked at variables available in the survey. It did not look at other potential mediating variables that may affect mental health. Researchers will want to conduct further research to determine if there are any other significant factors that play a role in mental health and compensation.

This data further supports the need for gender parity in compensation. Practitioners should look for ways of correcting salary disparities between genders.

6. ACKNOWLEDGEMENTS

We would like to thank and acknowledge the People's United Center for Women & Business at Quinnipiac University for helping to sponsor our efforts in this project.

7. REFERENCES

Adams-Prassl, A., Boneva, T., Golin, M., Rauh, C., (2022). The impact of the coronavirus lockdown on mental health: evidence from the United States. *Economic Policy*, 37(109), 139-155.

Araya R, Lewis G, Rojas G, & Fritsch R. (2003) Education and income: which is more important for mental health? *J Epidemiol Community Health.* 57(7), 501-505.

Asaduzzaman, M., Mashiyat, A. S., Roy, C. K., & Schneider, K. A. (2013, May). Answering questions about unanswered questions of stack overflow. In 2013 10th Working Conference on Mining Software Repositories (MSR) (pp. 97-100).

Barua, A., Thomas, S. W., & Hassan, A. E. (2014). What are developers talking about? an analysis of topics and trends in stack overflow. *Empirical Software Engineering*, 19(3), 619-654.

Bernstein G. & Shaw K. (1997) Practice parameters for the assessment and treatment of children and adolescents with anxiety disorders. *Journal of the American Academy of Child and Adolescent Psychiatry* 36, 69S–84S.

Center for Disease Control (2022) About Mental Illness retrieved May 21, 2022 from https://www.cdc.gov/mentalhealth/learn/.

Cochrane R, Stopes-Roe M. (1981) Women, marriage, employment and mental health*. Br J Psychiatry*. 139(5), 373–81.

Diperna J. & Elliott S. (2002) Promoting academic enablers to improve student achievement: an introduction to the mini-series. *School Psychology Review* 31, 293–297.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquez, J., Gruber, B., Lafourcade, B., Leitao, P., Munkemuller, T., McClean, C., Osborne, P., Reineking, B., Schroder, B., Skidmore, A., Zurell, D., & Lautenbach, S. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1), 27-46.

Gresenz, C. R., Sturm, R., and Tang, L. (2001). Income and mental health: unraveling community and individual level relationships. *The Journal of Mental Health Policy and Economics*. 4, 197-203.

Gutman L., Sameroff A., & Cole R. (2003). Academic growth curve trajectories from 1st grade to 12th grade: effects of multiple social risk factors and preschool child factors. *Developmental Psychology,* 39, 777–790.

Hylton, S., Ice, L., & Krutsch, E. (2022). What the long-term impacts of the COVID-19 pandemic could mean for the future of IT jobs retrieved May 30, 2022 from https://www.bls.gov/opub/btn/volume-11/what-the-long-term-impacts-of-the-covid-19-pandemic-could-mean-for-the-future-of-it-jobs.htm

Kessler, R. C., Foster, C. L., Saunders, W. B., and Stang, P.E. (1995). Social Consequences of Psychiatric Disorders, I: Educational Attainment." *American Journal of Psychiatry* 152(7), 1026-1032.

Kose, T. (2020) Gender, Income and Mental Health: The Turkish Case. PLoS One. 2020 Apr 29; 15(4):e0232344. doi: 10.1371/journal.pone.0232344. PMID: 32348361; PMCID: PMC7190175.

Lantz, P., House, J., Mero, R., & Williams, D. (2005). Stress, Life Events, and Socioeconomic Disparities in Health: Results from the Americans' Changing Lives Study. *Journal of Health & Social Behavior*, 46(3), 274-288.

Lewis G., Bebbington P., Brugha T., Farrell, M., Gill, B., Jenkins, R., Meltzer, H. (1998) Socio-economic status, standard of living and neurotic disorder. *Lancet,* 352, 605–609.

Li, X., Lu, P., Hu, L., Huang, T., and Lu, L. (2020). Factors associated with mental health results among workers with income losses exposed to COVID-19 in China. *International Journal of Environmental Research and Public Health*. 17(15), https://doi.org/10.3390/ijerph17155627

Lindau, S. T., Makelarski, J. A., Boyd, K., Doyle, K. E., Haider, S., Kumar, S., Lee, N. K., Pinkerton, E., Tobin, M., Vu, M., Wroblewski, K. E., & Lengyel, E. (2021). Change in health-related socioeconomic risk factors and mental health during the early phase of the COVID-19 pandemic: a national survey of US women. *Journal of Women's Health*, 30(4), 502-513.

Mental Health America, Inc. (2022). The State Of Mental Health In America retrieved May 18, 2022 from https://mhanational.org/issues/state-mental-health-america

Mayo Clinic (2022a). Mental Illness retrieved May 18, 2022 from https://www.mayoclinic.org/diseases-conditions/mental-illness/symptoms-causes/syc-20374968

Mayo Clinic (2022b). Depression in Women: Understanding the Gender Gap retrieved May 30, 2022 from https://www.mayoclinic.org/diseases-conditions/depression/in-depth/depression/art-20047725

McLeod, J., Uemura, R. & Rohrman, S. (2012). Adolescent Mental Health, Behaviour Problems, and Academic Achievement, *Journal of Health & Social Behavior,* 53(4), 482-537.

Mirowsky, J., & Ross, C. (2003). Education, Social Status, and Health. Hawthorne, NY, Aldine de Gruyter.

National Institute of Mental Health (2022). Mental Illness retrieved May 21, 2022 from https://www.nimh.nih.gov/health/statistics/mental-illness

Needham, B„ Crosnoe, R. & Müller, C. (2004). Academic Failure in Secondary School: The Inter-Related Role of Health Problems and Educational Context. *Social Problems,* 51(4), 569—586.

OSMI (2016). OSMI Mental Health in Tech Survey 2016 retrieved May 30, 2022 from https://osmi.typeform.com/report/Ao6BTw/U76z

Paula da Silva Ramos, A., Fernandes de Souza Ribeiro, J., Lima Trajano, E. T., Aurélio Dos Santos Silva, M., & Alexsandra da Silva Neto Trajano, L. (2022). Hospitalized Women Have Anxiety and Worse Mental Health Scores than Men. Psychological reports, 332941221088967. Advance online publication. https://doi.org/10.1177/00332941221088967

Pieh, C., Budimir, S., and Probst, T. (2020). The effect of age, gender, income, work, and physical activity on mental health during coronavirus disease (COVID-19) lockdown in Austria. *Journal of Psychosomatic Research*. 136, https://doi.org/10.1016/j.jpsychores.2020.110186

Pieh, C., Budimir, S., Delgadillo, J., Barkham, M., Fontaine, J. R. J., and Probst, T. (2021). Mental health during COVID-19 lockdown in the United Kingdom. *Psychosomatic Medicine*. 83(4), 328-337.

Prowse, R., Sherratt, F., Abizaid, A., Gabrys, R. L., Hellemans, K. G. C., Patterson, Z. R., and McQuaid, R. J. (2021). Coping with the COVID-19 pandemic: examining gender differences in stress and mental health among university students. *Frontiers in Psychiatry*. 12, https://doi.org/10.3389/fpsyt.2021.650759

Purtle, J. (2020). COVID-19 and mental health equity in the United States. *Social Psychiatry and Psychiatric Epidemiology*. 55, 969-971.

Reid, R., Gonzalez J., Nordness P., Trout, A. and Epstein, M. (2004), A meta-analysis of the academic status of students with emotional/behavioral disturbance. *Journal of Special Education,* 38, 130–143.

Ritsher, J., Warner, V., Johnson J., & Dohrenwend, B. (2001) Inter-Generational Longitudinal Study of Social Class and Depression: A Test of Social Causation and Social Selection Models." *British Journal of Psychiatry,* 178(40), 84-90.

Schieman, S., & Plickert. G. (2008), How Knowledge Is Power: Education and the Sense of Control. *Social Forces,* 87(1), 153-183

Seedat S, Scott KM, Angermeyer MC, Berglund P, Bromet EJ, Brugha TS, et al. (2009). Cross-national associations between gender and mental disorders in the World Health Organization World Mental Health Surveys. *Arch Gen Psychiatry*, 66(7), 785–795.

Simon RW. (1995). Gender, multiple roles, role meaning, and mental health. *J Health Soc Behav*, 36(2), 182–194.

SingleCare (2022). Mental Health Statistics 2022, America retrieved May 18, 2022 from https://www.singlecare.com/blog/news/mental-health-statistics/

Stack Overflow retrieved May 18, 2022 from https://insights.stackoverflow.com/survey/2020#overview

Strohschein, L. (2005). Household income histories and child mental health trajectories. *Journal of Health and Social Behavior*. 46(4), https://doi.org/10.1177%2F002214650504600404

Thompson, D. (2022). Mental Health – An Important Conversation in the Tech Industry retrieved May 30, 2022 from https://www.techtimes.com/articles/271446/20220204/mental-health-an-important-conversation-in-the-tech-industry.htm#:~:text=Bhateja%20added%2C%20%22The%20high%20stress,with%20a%20mental%20health%20condition.

Treude, C., & Robillard, M. P. (2016, May). Augmenting api documentation with insights from stack overflow. In 2016 *IEEE/ACM 38th International Conference on Software Engineering (ICSE)* (pp. 392-403).

US Bureau of Labor Statistics, (2021). Unemployment rises in 2020, as the country battles the COVID-19 pandemic retrieved July 29, 2022 from https://www.bls.gov/opub/mlr/2021/article/unemployment-rises-in-2020-as-the-country-battles-the-covid-19-pandemic.htm

Weich S, & Lewis G. (1998). Poverty, unemployment and the common mental disorders: a population-based cohort study. *BMJ,* 317, 115–119.

Wildman, J. (2003). Income related inequalities in mental health in Great Britain: analyzing the causes of health inequality over time. *Journal of Health Economics*. 22(2), 295-312.

World Health Organization (2022) COVID-19 pandemic triggers 25% increase in prevalence of anxiety and depression worldwide, retrieved May 18, 2022 from https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide.

Xie, Y., Xu, E., & Al-Aly, Z. (2022). Risks of mental health outcomes in people with covid-19: cohort study. *BMJ*, 376.

Appendices

MentalHealth	Mean	N
▪ concentration and/or memory disorder (e.g. ADHD)	$130,641	628
▪ concentration and/or memory disorder (e.g. ADHD); ▪ mood or emotional disorder (e.g. depression, bipolar disorder)	139,442	139
▪ concentration and/or memory disorder (e.g. ADHD); ▪ mood or emotional disorder (e.g. depression, bipolar disorder); ▪ anxiety disorder	122,269	250
▪ concentration and/or memory disorder (e.g. ADHD); ▪ mood or emotional disorder (e.g. depression, bipolar, disorder) ▪ anxiety disorder; ▪ Autism / an autism spectrum disorder (e.g. Asperger's)	132,636	55
▪ concentration and/or memory disorder (e.g. ADHD); ▪ mood or emotional disorder (e.g. depression, bipolar disorder); ▪ anxiety disorder; ▪ Autism / an autism spectrum disorder (e.g. Asperger's); ▪ Or, in your own words:	80,000	1
▪ concentration and/or memory disorder (e.g. ADHD); ▪ mood or emotional disorder (e.g. depression, bipolar disorder); ▪ anxiety disorder; ▪ Or in your own words:	98,800	6
▪ concentration and/or memory disorder (e.g. ADHD); ▪ mood or emotional disorder (e.g. depression, bipolar disorder); ▪ Autism / an autism spectrum disorder (e.g. Asperger's)	137,263	19
▪ concentration and/or memory disorder (e.g. ADHD); ▪ mood or emotional disorder (e.g. depression, bipolar disorder); ▪ Or, in your own words:	140,333	6
▪ concentration and/or memory disorder (e.g. ADHD); ▪ anxiety disorder	115,612	129
▪ concentration and/or memory disorder (e.g. ADHD); ▪ anxiety disorder; ▪ Autism / an autism spectrum disorder (e.g. Asperger's)	125,461	21
▪ concentration and/or memory disorder (e.g. ADHD); ▪ anxiety disorder; ▪ Or, in your own words:	107,500	2
▪ concentration and/or memory disorder (e.g. ADHD); ▪ Autism / an autism spectrum disorder (e.g. Asperger's)	132,129	54
▪ concentration and/or memory disorder (e.g. ADHD); ▪ Or, in your own words:	117,133	14

▪ mood or emotional disorder (e.g. depression, bipolar disorder)	126,893	317
▪ mood or emotional disorder (e.g. depression, bipolar disorder); ▪ anxiety disorder	124,327	301
▪ mood or emotional disorder (e.g. depression, bipolar disorder); ▪ anxiety disorder; ▪ Autism / an autism spectrum disorder (e.g. Asperger's)	113,117	25
▪ mood or emotional disorder (e.g. depression, bipolar disorder); ▪ anxiety disorder; ▪ Or, in your own words:	124,850	2
▪ mood or emotional disorder (e.g. depression, bipolar disorder); ▪ Autism / an autism spectrum disorder (e.g. Asperger's)	165,600	18
▪ mood or emotional disorder (e.g. depression, bipolar disorder); ▪ Or, in your own words:	159,625	8
▪ anxiety disorder	131,383	379
▪ anxiety disorder; ▪ Autism / an autism spectrum disorder (e.g. Asperger's)	130,633	24
▪ Anxiety disorder; ▪ Autism / an autism spectrum disorder (e.g. Asperger's); ▪ Or, in your own words:	93,000	2
▪ anxiety disorder; ▪ Or, in your own words:	101,400	10
▪ Autism / an autism spectrum disorder (e.g. Asperger's)	142,353	109
▪ Autism / an autism spectrum disorder (e.g. Asperger's); ▪ Or, in your own words:	250,000	1
None of the above	139,548	5,011

Appendix B Regression and Collinearity Results

Coefficients[a]

Model		Unstandardized Coefficients		Standard. Coeff.			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	64029.663	4812.012		13.306	.000		
	Mental HealthN	-5143.495	1656.803	-.035	-3.104	.002	.960	1.042
	Ethnicity	5917.824	1861.583	.036	3.179	.001	.982	1.018
	EducN	6306.749	746.751	.095	8.446	.000	.976	1.025
	Woman	-21409.143	3217.936	-.075	-6.653	.000	.980	1.020
	OtherN	-13370.434	10438.117	-.014	-1.281	.200	.995	1.005
	AgeN	14390.041	734.509	.220	19.591	.000	.981	1.019

a. Dependent Variable: Compensation,

Collinearity Diagnostics[a]

Model	Dim.	Eigen Value	Condition Index	Variance Proportions (Constant)	Mental HealthN	Ethnic,	EducN	Woman	OtherN	AgeN
1	1	4.245	1.000	.00	.02	.01	.00	.01	.00	.01
	2	1.002	2.059	.00	.00	.00	.00	.05	.93	.00
	3	.925	2.142	.00	.02	.00	.00	.87	.04	.00
	4	.608	2.643	.00	.90	.00	.00	.07	.02	.01
	5	.138	5.547	.00	.00	.26	.00	.01	.00	.64
	6	.065	8.095	.03	.00	.59	.28	.00	.00	.24
	7	.018	15.573	.97	.05	.14	.71	.00	.00	.11

a. Dependent Variable: Compensation

Measuring Learners' Cognitive Load when Engaged with an Algorithm Visualization Tool

Razieh Fathi
rfathi@smith.edu
Department of Computer Science
Smith College
Northampton, MA 01060 USA

James D. Teresco
jteresco@siena.edu
Department of Computer Science
Siena College
Loudonville, NY 12211 USA

Kenneth Regan
regan@buffalo.edu
Department of Computer Science
University at Buffalo
Buffalo, NY 14260 USA

Abstract

We present results of a preliminary study that applies cognitive load theory (CLT) to investigate how students with different amounts of prior experience learn algorithms. We test the following assertions from the CLT framework: The high CL of algorithm learning comes from intrinsic CL – meaning the complexity of the information being processed. There is also high germane CL – that induced by the instructional intervention – in tasks designed to assess the learned knowledge. Lowering either of these two CLs results in measurable learning gains. Lowering the complexity of incremental steps is the key determinant of success. We investigated the extent to which students' previous knowledge and experience influence the process of learning algorithms. This also involved testing whether an algorithm visualization tool (Map-based Educational Tools for Algorithm Learning, METAL) improves the understanding of graph algorithms. Our study adapted an existing survey instrument developed by Klepsch, et al., to algorithmic thinking tasks and used it as a tool to measure CL components. We explored and measured three types of CL for breadth-first and depth-first graph traversal algorithms, and among three groups of participants, non-Computer Science students, beginning CS students, and more advanced CS students. Results include: (i) Among different types of CL, germane load was the most substantial type for all groups. Students with more background in CS showed lower levels of all types of CL. (ii) The three groups showed similar relative effects of intrinsic, germane, and extraneous CL. We discuss future research and limitations of the study.

Keywords: algorithm visualization, student engagement, cognitive load, student learning

1. INTRODUCTION

As computer science has opened wide to students of diverse backgrounds and different levels of prior experience, the educational community needs progressively better understanding of how to optimize learning across this spectrum. Algorithm visualization (AV) tools have been shown to be effective (Hansen et al. 2002) but they fit a wider picture. We gain insight by conducting an experiment that employs Cognitive Load Theory (CLT) (Paas et al. 2003a; Sweller et al. 2011) to see how students at various levels learn when guided by an AV tool.

The experiment involves university students at three levels of experience with computing: non-computer science majors, those early in the CS major, and those at a more advanced stage of the major. We first describe CLT and how it applies in this context. Then we present results that accord with expectations from previous work in CLT and draw further conclusions.

2. COGNITIVE LOAD THEORY

CLT aims to structure the analysis of what is commonly called the "learning curve." How can we define and measure the effort required to learn concepts? We want to measure the efficiency of a learning process as the proportion that drives acquired knowledge and skills, versus the part expended on incidentals of the learning process. This can inform the design and ordering of instructional materials, and also evaluate the efficacy of automated learning tools. Much of the effort goes into memorizing and retention, which mean both the memory immediately needed to function and the memory of how concepts and procedures are ordered so they can be efficiently recovered from notes.

According to CLT, cognitive workload is the level of measurable mental effort put forth by an individual in response to one or more cognitive tasks (Van Gog & Paas 2008). In other words, cognitive load can be defined as the ratio between the workload that directly leads to the acquisition of knowledge and skills, and the workload that is expended on incidentals of the learning process. In general, there are three categories of cognitive workload (Sweller et al. 2019), reflected also in (Klepsch et al. 2017; Klepsch & Seufert 2020):

- Intrinsic cognitive load (ICL)
- Extraneous cognitive load (ECL)
- Germane cognitive load (GCL)

Intrinsic cognitive load refers to the complexity of the information being processed. It also relates to the concept of element interactivity. Interactive elements have to be processed simultaneously in working memory for learning to begin. Consequently, learning new material with a high number of interacting elements will impose a high cognitive workload. The other two kinds of loads are not considered inherent to learning the material, but as imposed by the design of instructional units for that material. When the load imposed by the design is ineffective or detrimental for learning, it is called extraneous cognitive load; when it is effective for learning it is referred to as germane cognitive load (Sweller et al. 2019).

The overall key to improving learning for novices is reducing the undesirable parts of the cognitive load to allow maximum memory usage for learning (Morrison et al. 2016). One of the original assumptions of CLT is that the three basic types of load (ICL, ECL, and GCL) are additive (Paas et al. 2003); thus, if the ECL is using the capacity of working memory, little can be devoted to the GCL. Because working memory is considered to be a fixed size (Miller 1956), it falls upon the instructional designer to minimize the ECL, design appropriately for the ICL, and emphasize the GCL. To accomplish this, one must be able to measure the specific load components for any pedagogical intervention.

Context and Relevant Work

Sweller (Sweller 1988) proposed CLT, which articulated the association between cognitive resources and task demands in creating cognitive load. Key elements defined in (Groth-Marnat & Wright 2016) and (Van Gog & Paas 2008) are *schemata* and *schemas*. The former means cognitive structures representing generic knowledge, i.e., structures that do not contain information about particular entities, instances or events, but rather about their general form. People use schemata to organize current knowledge and provide a framework for future understanding. Examples of schemata include academic rubrics. Schemas, on the other hand, are single information elements that combine to form schemata.

Learning is considered to happen through schema construction, elaboration, and automation. Automation means execution without controlled processing through intensive and consistent practice (Van Gog & Paas 2008). Cognitive load is metered by resources that learners consume while performing tasks. In this model, working memory is a cognitive resource, but is a limited one; only a small fraction of elements can be consciously handled per unit time, especially

when they are novel or unfamiliar. However, long-term memory provides the ability to overcome the limitation of working memory, with the help of schemas (Xie et al. 2017).

In the field of educational research, CLT is mainly used to explain the effects of various forms of instructional design (Sweller et al. 2011). According to this theory, ICL is not directly affected by instructional design. It is related to element interactivity in learning materials and learners' prior knowledge. The level of ICL of a specific task is usually treated as depending on the level of element interactivity (Xie et al. 2017). An element can be anything that will be or has been presented, for example a concept or a procedure. Instructional materials with low element interactivity allow single (or several) element(s) to be processed with little or even no reference to other elements, thus resulting in a low ICL; however, high element interactivity materials contain elements that heavily interact with each other and cannot be processed separately, leading to a high ICL. The theory supports the position that GCL is directly beneficial to learning, whereas ECL only is detrimental to learning. In particular, GCL is imposed by cognitive processes of active schema construction, such as clarifying, inferring, and organizing, whereas ECL obstructs schema construction and automation.

The total cognitive load during information processing is the sum of the three kinds of cognitive loads. One important objective of instructional design is to ensure that the total cognitive load is within the learner's cognitive capacity, in order to avoid cognitive overload (Paas et al. 2003b). Techniques used to measure cognitive load include subjective rating scales, dual-task performance, and physiological measures (Antonenko et al. 2010; Paas et al. 2003b; Whelan 2007). Paas (Paas 1992) introduced the mental effort scale, which was a modified version of Bratfisch, Borg, and Dornic's scale (Bratfisch et al. 1972) for measuring perceived task difficulty. Paas's 9-point mental effort scale included one item that asked learners to report how much mental effort they invested when learning the material. Since then, the mental effort or perceived difficulty scale has been widely used in research in the field of learning and instruction because it is easy to administer, is non-invasive, and has good reliability and validity (Paas et al. 2003b)

3. USING METAL AS AN INTERACTIVE ALGORITHM VISUALIZATION

Transfer of learning occurs when people apply information, strategies, and skills they have learned to a new situation or context (Olson 2015). Transferring this learning performance generally requires additional instructional support that allows learners to go below the schema level in the level of hierarchy of learning and to understand the rationale of individual solution steps. One promising avenue to support learners in this type of reasoning is to embed interactive visualizations within an example-based hypermedia environment (Van Merriënboer et al. 2003). This study uses a variant of this idea, based on an interactive AV tool. The AV system provided by the Map-based Educational Tools for Algorithmic Learning (METAL) project (Teresco et al. 2018) has several advantages: scalability, a customizable API, visualizations that show the progress of algorithms overlaid on Leaflet Maps, color-coded tables showing contents of data structures, and example real-world data sets in a variety of sizes. These all enhance student engagement (Teresco et al. 2018). Figures 1 and 2 show a snapshot of METAL's AV system in action for the two algorithms used in our study: breadth-first search within a graph (BFS) and depth-first search within a graph (DFS).

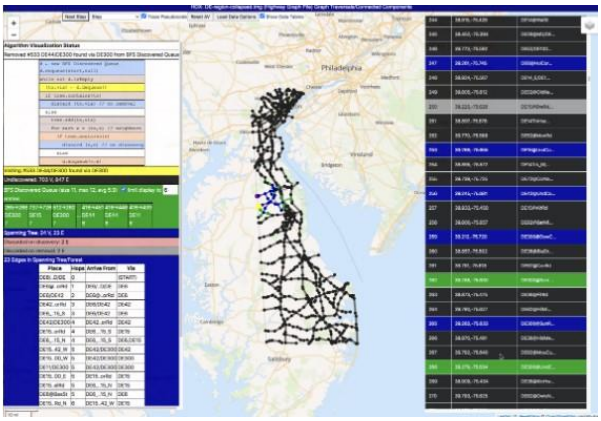


Fig. 1. A METAL AV in progress using the BFS algorithm on the Delaware region graph. The violet dot shows the starting vertex. Blue vertices and edges have been found to be part of the spanning tree. Green vertices and edges are candidates have been "discovered" and are in the queue of candidates to be added to the spanning tree in subsequent steps. The yellow edge and vertex just came out of the discovered queue as the next candidate to be added to the spanning tree.

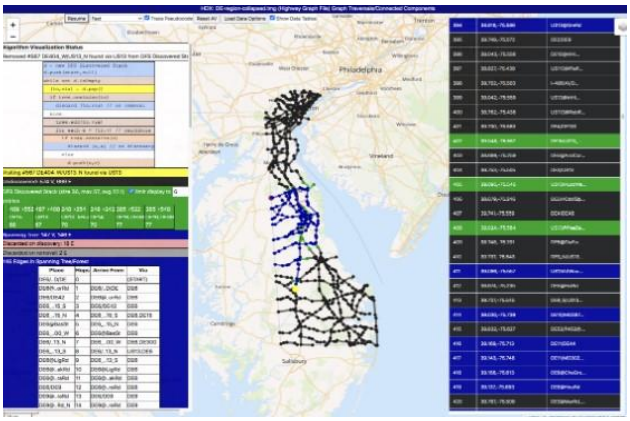


Fig. 2. A METAL AV in progress using the DFS algorithm on the Delaware region graph. Colors here match those described in Figure 1, except here the green discovered vertices and edges are stored in a stack rather than a queue.

4. METHOD

Our cognitive load measurement survey is adapted from Klepsch, et al (Klepsch et al. 2017; Klepsch & Seufert 2020). It includes elements to measure each of the three cognitive load components (English translations).

- For this task, many things needed to be kept in mind simultaneously (ICL).
- This task was very complex (ICL).
- For this task, I had to highly engage myself (GCL).
- For this task, I had to think intensively what things meant (GCL).
- During this task, it was exhausting to find the important information (ECL).
- The design of this task was very inconvenient for learning (ECL).
- During this task, it was difficult to recognize and link the crucial information (ECL).

Participants responded to each item using a Likert score from 0 ("absolutely wrong") to 7 ("absolutely right"). For the study herein, the questions were as follows:

- For BFS/DFS, many things needed to be kept in mind simultaneously (ICL).
- BFS/DFS was very complex (ICL).
- I made an effort, not only to understand several details, but to understand the overall context (GCL).
- My point while dealing with BFS/DFS was to understand everything correctly (GCL).
- The learning task consisted of elements supporting my comprehension of the task (GCL).
- During this task, it was exhausting to find the important information (ECL).
- The design of this task was very inconvenient for learning (ECL).
- During this task, it was difficult to recognize and link the crucial information (ECL).

Study Participants

Participants were undergraduate students, aged 18-24, in a university in the United States. Participants included both Computer Science (CS) majors, and students majoring in other disciplines. The CS majors were furthered divided into those at the CS1 stage of computer science and those at CS2 or higher (by the ACM classification). Thus, our participants are divided into 3 groups of 15 students each: (i) Non-CS majors (denoted hereafter as "NCS"), (ii) CS1 students (denoted as "CS1"), and (iii) CS2 or higher level students (denoted as "CS2+"). The cognitive load survey was administered after students were exposed to learning tasks and interview questions. Surveys were online and took 10 minutes to complete for each task. During data collection, cognitive load surveys were monitored in a Zoom meeting. No invalid surveys were returned. A total of 90 cognitive load surveys, 45 each for BFS and for DFS were collected.

Design of the Study

The study involves two families of algorithms: breadth-first search (BFS) and depth-first search (DFS). Both BFS and DFS are accessible at some level to both an advanced CS major and a non-major. The procedure of the study is divided into two blocks. One for BFS and one for DFS. After completing informed consent, the process below was used first for the BFS algorithm, then repeated for the DFS algorithm.

(1) The participant was asked several interview questions relevant to the algorithm (CS1 and CS2+ only, as NCS participants are assumed to be unfamiliar with the algorithms).

(2) The participant watched the METAL AV tutorial video for the algorithm.

(3) The participant uses the interactive METAL AV for the algorithm.

(4) The participant was asked several knowledge questions with different levels of complexity regarding the interaction with METAL and knowledge related to the algorithm.

(5) The participant completes the cognitive load survey.

5. RESULTS AND DISCUSSIONS

The experimental design gives several axes for comparisons. First, we compare the cognitive load experienced in the ICL, GCL and and ECL categories, which are reflected by three groups of questions: Q1-Q2, Q3-Q5, and Q6-Q8. Then we compare the three groups of non-majors, CS1, and CS2, repeating for BFS (Figures 3-5) and DFS (Figures 6-8). Finally we compare experiences with BFS vs. DFS holding other factors the same (Figures 9-11).

In these comparisons, we try to refute a null hypothesis expressing that there is no difference between groups, i.e., that the inputs from the groups come from one underlying distribution. The basic Analysis of Variance (ANOVA) test presumes that this distribution is normal. For one student, the responses are drawn from the 0 to 7 Likert scale. We avoid the controversial presumption that each individual student's responses can be treated as drawn from a normal distribution centered somewhere on the Likert scale. Instead, for each of the eight questions, and within each group of 15 students, we average the responses to the question and input the mean as one data item. By appeal to the Central Limit Theorem, these means are representative of a normal distribution. As also stated in the footnoted excerpt from (Willett nd), we are shy of the conventional 30 for such appeal, but we compensate by having three groups of 15 and by the observation that our individual Likert responses do not have extreme polarization—that is, they do not have two modes on the 0-7 scale where some students strongly agree and others strongly disagree. We still find significance even with just 8 data points per item in the ANOVA, while our results are conservative compared to the alternative procedure of treating individual responses as normally distributed. We plot sums out of 105 rather than averages out of 7 for each group; this makes no difference to the ANOVA.

The first null hypothesis (NH) we try to refute is that there is no difference in the students' experience of cognitive load between the items identified as ICL, GCL, and ECL. Under NH, we would be supposing that the sampled questions are indistinguishable from randomly drawn responses about stages of cognitive load. As per the original design in Klepsch, et al., there is sufficient homogeneity in what each of Q1,...,Q8 addresses—this also reflects the intent to divide the learning exercise into "steps" that are reasonably uniform.

BFS Algorithm Studies

Participants in all groups first worked with the BFS algorithm and completed the surveys.

Non-CS Majors. Recall that Q1 and Q2 are intended to measure ICL; Q3-Q5 measure GCL and Q6-Q8 measure ECL.

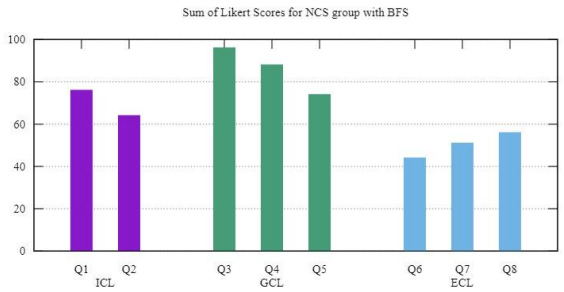


Fig. 3. Sum of Likert scores by question for the BFS algorithm study's 15 NCS participants.

In Figure 3, we see the GCL is highest, followed by ICL, with ECL the lowest. To test if there is a significant difference among these three, we use an ANOVA analysis to obtain a probability (p-value). This was computed as 0.012, less than a significance threshold of 0.05, indicating a significant difference. Details of this analysis are shown in Table 1.

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1913.208	2	956.604	12.181	.012
Within Groups	392.667	5	78.533		
Total	2305.875	7			

Table 1. Results of the ANOVA analysis for the BFS algorithm study's NCS participants. "Groups" here are the three types of cognitive load measured.

We follow this by using Tukey's HSD test (Tukey 1949) post hoc to indicate which groups in the sample differ. Tukey's test uses the honest significant difference, a number that represents the distance between groups, to compare every mean with every other mean. The results of this test indicate a significant difference between GCL and ECL. For an example of this, we quote Willett (Willett nd) from the Simulation Canada website: "[It] is common to see ordinal data analyzed using parametric tests, such as the t-test or an ANOVA. Sometimes this is appropriate and sometimes it is not. So when can parametric tests, which are generally more sensitive and more powerful, be used? Only when the ordinal data meets all of the assumptions of the parametric test. These are: 1. The sampling distribution (not necessarily the data itself) is normally distributed. This will be true if 1. Sample

size (n) is greater than 30; or 2. n<30 and the data appears to be normally distributed on inspection."

CS1 students. Figure 4 shows that the three questions that measure GCL have the highest individual cognitive load for participants, and account more than half the total cognitive load.

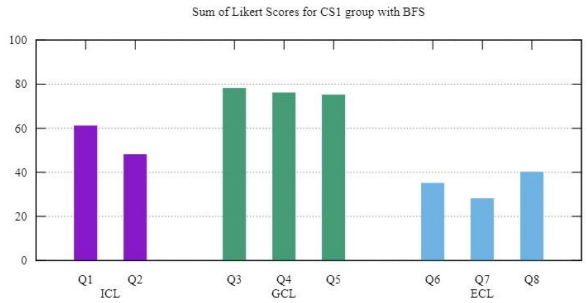


Fig. 4. Sum of Likert scores by question for the BFS algorithm study's 15 CS1 participants.

As expected, ECL is the lowest. The BFS algorithm is new to this group of participants, but there is a difference in learning BFS between CS1 and NCS. CS1 students have some familiarity with the queue data structure at the core of the BFS algorithm, so we would expect this group to have some better schema creation, resulting in lower GCL, based on this previous experience that NCS participants did not have. This is observed in the GCL questions. ICL and ECL are also lower for the CS1 group compared to the NCS group.

For this group, ANOVA analysis (Table 2) gives a p-value of 0.001, again below the significance threshold of 0.05, indicating significant differences among the three types of cognitive load. The post hoc Tukey test also indicates significant differences between each pair of cognitive load types.

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2647.042	2	1323.521	40.891	.001
Within Groups	161.833	5	32.367		
Total	2808.875	7			

Table 2. Results of the ANOVA analysis for the BFS algorithm study's CS1 participants. "Groups" here are the three types of cognitive load measured.

CS2+ Students. Similarly to the NCS and CS1 groups, Figure 5 shows that the GCL is the largest component of cognitive load among the CS2+ group. CS2+ participants were familiar with the BFS algorithm but had not seen it recently (based on their pre-test interview) and had studied relevant data structures. The BFS AV for these students was more of a refresher. There was no significant difference between CS1 and CS2+ participants' cognitive loads.

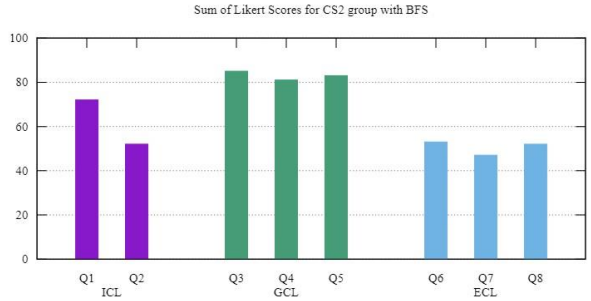


Fig. 5. Sum of Likert scores by question for the BFS algorithm study's 15 CS2+ participants.

The ANOVA analysis for CS2+ (Table 3) shows a significant overall difference among three cognitive loads (p-value 0.006). The post hoc Tukey test indicates a significant difference between ICL and GCL, and between ECL and GCL, but not between ICL and ECL.

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1603.208	2	801.604	17.528	.006
Within Groups	228.667	5	45.733		
Total	1831.875	7			

Table 3. Results of the ANOVA analysis for the BFS algorithm study's CS2+ participants. "Groups" here are the three types of cognitive load measured.

DFS Algorithm Studies

Participants in all groups next worked with the DFS algorithm and completed the surveys

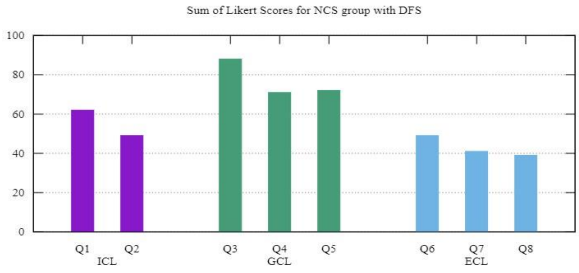


Fig. 6. Sum of Likert scores by question for the DFS algorithm study's 15 NCS participants.

Non-CS Majors. Results for the NCS group's interaction with DFS in Figure 6 show that all types of cognitive load are lower than for the same group with BFS, indicating a lower level of

difficulty (unsurprising, since they had done BFS first). As was the case with BFS for this group, GCL is higher than the other loads for DFS.

NCS participants had a significant difference among the three types of load, as indicated by the p-value 0.009 obtained from the ANOVA analysis (Table 4). The post hoc test shows there is a significant difference only between ECL and GCL.

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1764.375	2	882.188	13.677	.009
Within Groups	322.500	5	64.500		
Total	2086.875	7			

Table 4. Results of the ANOVA analysis for the DFS algorithm study's NCS participants. "Groups" here are the three types of cognitive load measured.

CS1 Students. With DFS, the CS1 group (Figure 7) GCL is higher than ICL and ECL.

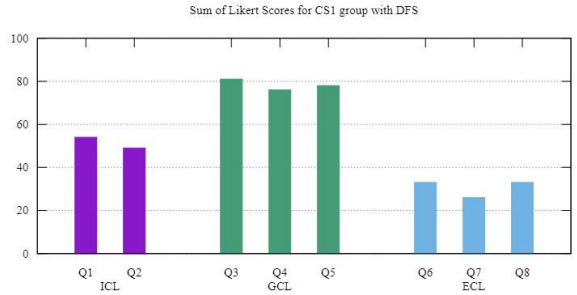


Fig. 7. Sum of Likert scores by question for the DFS algorithm study's 15 CS1 participants.

These results are very similar to this group's results with BFS in all three types of cognitive load, and indicate that the DFS and BFS algorithm for CS1 participants were about the same level of difficulty.

The ANOVA analysis here (Table 5) gives a p-value 0.001, demonstrating significant differences among the three types of cognitive load. The post hoc Tukey test shows there is a significant difference between each pair of ICL, ECL, and GCL.

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3421.667	2	1710.833	147.911	<.001
Within Groups	57.833	5	11.567		
Total	3479.500	7			

Table 5. Results of the ANOVA analysis for the DFS algorithm study's CS1 participants. "Groups" here are the three types of cognitive load measured.

CS2+ Students. Again for DFS with the CS2+ group, Figure 8 shows that GCL is highest.

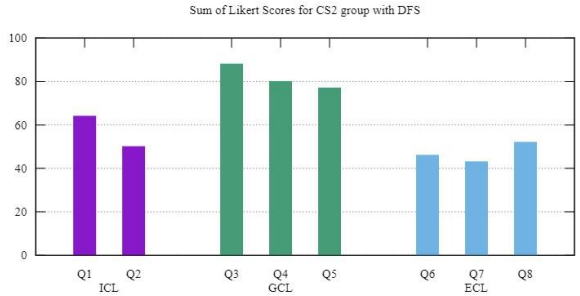


Fig. 8. Sum of Likert scores by question for the DFS algorithm study's 15 CS2+ participants.

ANOVA analysis (Table 6) gives a significant difference among the types of load (p-value 0.003). The post hoc Tukey test shows there is a significant difference between GCL and ECL, and between GCL and ICL, but not between ICL and ECL. CS2+ participants have the most familiarity with the algorithm and have little need for schema creation to store new knowledge, meaning less difference between ICL and ECL.

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1883.333	2	941.667	23.005	.003
Within Groups	204.667	5	40.933		
Total	2088.000	7			

Table 6. Results of the ANOVA analysis for the DFS algorithm study's CS2+ participants. "Groups" here are the three types of cognitive load measured.

Comparisons Between BFS and DFS

Figures 9 (for the NCS group), 10 (for the CS1 group), and 11 (for the CS2+ group), show side-by-side comparisons of the results for BFS and DFS surveys presented earlier in this section. For NCS participants, we see that the cognitive loads are smaller for DFS than for BFS. For the CS1 and CS2+ groups, we observe a reduction in ICL and ECL for DFS compared to BFS.

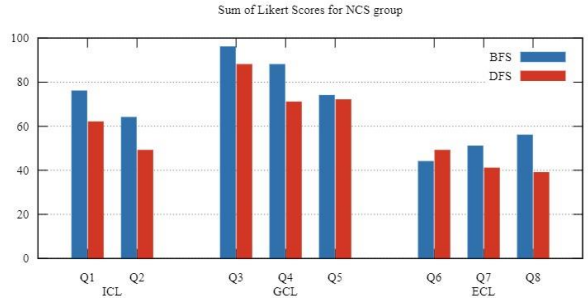


Fig. 9. Side-by-side comparison of the sum of Likert scores by question for the BFS algorithm study and DFS algorithm study's 15 NCS participants.

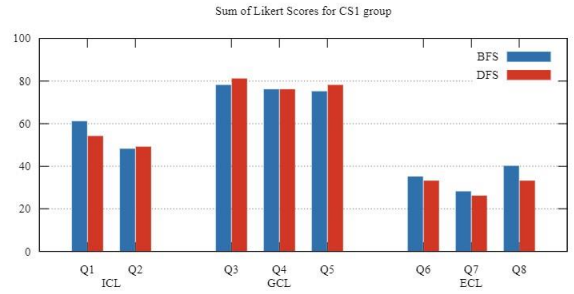


Fig. 10. Side-by-side comparison of the sum of Likert scores by question for the BFS algorithm study and DFS algorithm study's 15 CS1 participants.

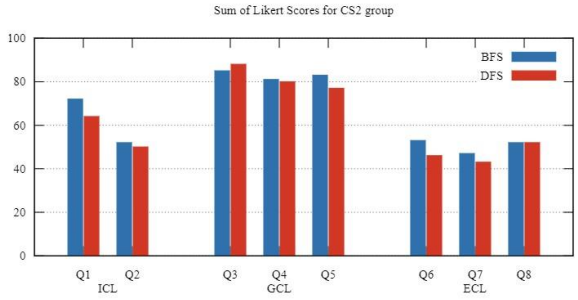


Fig. 11. Side-by-side comparison of the sum of Likert scores by question for the BFS algorithm study and DFS algorithm study's 15 CS2+ participants.

Discussion

In summary, our BFS and DFS studies produced the following results:

- Students with more CS background (CS1 and CS2+) showed lower levels of all three types of cognitive load.
- Among types of cognitive load, GCL was the most substantial for all groups.
- There was less difference between the CS1 and CS2+ groups in regard to cognitive load compared to the NCS group.

Comparing the differences between the BFS and DFS algorithms, we obtained these results:

- There was not significant difference between the two algorithms for the groups that have a background in CS (CS1 and CS2+).
- There was a higher level of cognitive load for BFS than DFS for the NCS group.
- GCL was the highest type of cognitive load for both algorithms.

According to CLT literature, we can reduce ICL in two ways:

1. The segmenting principle (Mayer and Moreno 2010). The goal of this principle is to reduce element interactivity by presenting information step by step. This process helps learners without prior knowledge to organize the incoming information. The METAL AV user interface presents the algorithm in a step-by-step manner, possibly explaining the low ICL in spite of the highly intrinsic nature of algorithmic learning.
2. The pre-training principle (Mayer & Pilegard 2005). According to this principle, ICL is reduced by providing the learner with information about the content before starting with the learning material. Increasing the learner's prior knowledge supports the integration of new information. In the design of our study, students first watched a video of the algorithms to gain some familiarity with the topic. That might be another factor that helped to reduce ICL.

6. LIMITATIONS

Due to pandemic protocols, the algorithm learning experiment was conducted fully online through use of Zoom web conferencing. Participants were monitored during all the steps of study. If we could repeat the experiment in person, it is unclear if we would obtain similar results. Also, the motivation of our subjects to participate in the study is another key factor to consider. The background and previous knowledge of the population under study can affect the results. The study's population size is also a limiting factor. As discussed below, this experiment has the potential to be repeated with a larger sample size. The fact that DFS was learned before BFS is potentially a confounding factor. We are hoping to repeat the experiment with random order and then see the differences.

7. CONCLUSIONS AND FUTURE WORK

We hypothesized that the high level of cognitive load related to algorithm learning comes from ICL. We found that METAL has a positive impact on the learning process over all participants by helping to reduce ICL relative to GCL. We found a significant difference between ICL and GCL within each of the CS1 and CS2+ groups. This indicates that METAL was most effective for those

already with some CS background. This runs counter to intuition that visual tools may have greatest impact for neophytes and argues their aptness within core CS curricula.

We see this study as a first step that can enable many additional studies. The game plan for this includes the following:

(1) Find a more accurate and reliable measurement of cognitive load, especially as related to learning algorithms.

(2) Expand the use of the cognitive load survey in larger size algorithms classrooms.

(3) Short of being able to scale up the study in its entirety, we can consider partial questionnaires given to larger groups that can provide supplementary information relevant enough to buttress the conclusions.

(4) After the pandemic, it will be possible to employ physiological measurement tools such as eye-trackers to measure the cognitive process of algorithm learning. As a visual tool, METAL is suitable for this as well and will be a key component in this plan.

(5) Replicate the study with the different groups of students and changing the order in which BFS and DFS are introduced.

Points 1–3 raise the following general research question: Can we obtain results measuring cognitive load as accurately as the complex and time-consuming study that was used here with a streamlined study that can more reasonably be scaled to larger groups That could also mean we could do multiple studies or have multiple groups for a study within one larger cohort, gathering much more data in much less time.

8. REFERENCES

Antonenko, P., Paas, F., Grabner, R., & Van Gog, T. (2010). Using electroencephalography to measure cognitive load. Educational psychology review, 22(4), 425-438.

Bratfisch, O. (1972). Perceived Item-Difficulty in Three Tests of Intellectual Performance Capacity.

Groth-Marnat, G. (2009). Handbook of psychological assessment. John Wiley & Sons.

Hansen, S., Narayanan, N. H., & Hegarty, M. (2002). Designing educationally effective algorithm visualizations. Journal of Visual Languages & Computing, 13(3), 291-317.https://doi.org/10.1006/jvlc.2002.0236

Klepsch, M., Schmitz, F., & Seufert, T. (2017). Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. Frontiers in psychology, 8, 1997.https://doi.org/10.3389/fpsyg.2017.01997

Klepsch, M., & Seufert, T. (2020). Understanding instructional design effects by differentiated measurement of intrinsic, extraneous, and germane cognitive load. Instructional Science, 48(1), 45-77.

Mayer, R. E., & Moreno, R. E. (2010). Techniques that reduce extraneous cognitive load and manage intrinsic cognitive load during multimedia learning.

Mayer, R. E. (2005). The Cambridge Handbook of Multimedia Learning: Principles for Managing Essential Processing in Multimedia Learning: Segmenting, Pretraining, and Modality Principles.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological review, 63(2), 81.

Morrison, B. B., Margulieux, L. E., Ericson, B., & Guzdial, M. (2016, February). Subgoals help students solve Parsons problems. In Proceedings of the 47th ACM Technical Symposium on Computing Science Education (pp. 42-47).

Olson, M. H. (2015). Introduction to theories of learning. Routledge.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. Educational psychologist, 38(1), 1-4.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive science, 12(2), 257-285. John Sweller, Paul Ayres, and Slava Kalyuga. 2011. Measuring cognitive load. In Cognitive Load Theory. Springer, 71–85.

Sweller, J., van Merriënboer, J. J., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. Educational Psychology Review, 31(2), 261-292.

Teresco, J. D., Fathi, R., Ziarek, L., Bamundo, M.,

Pengu, A., & Tarbay, C. F. (2018, February). Map-based algorithm visualization with METAL highway data. In Proceedings of the 49th ACM Technical Symposium on Computer Science Education (pp. 550-555).

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. Biometrics, 99-114.

Van Gog, T., & Paas, F. (2008). Instructional efficiency: Revisiting the original construct in educational research. Educational psychologist, 43(1), 16-26.

Van Merriënboer, J. J., Kirschner, P. A., & Kester, L. (2003). Taking the load off a learner's mind: Instructional design for complex learning. Educational psychologist, 38(1), 5-13.

Whelan, R. R. (2007). Neuroimaging of cognitive load in instructional multimedia. Educational Research Review, 2(1), 1-12.

Timothy Willett. n.d.. Analyzing Likert Scale Data: The Rule of N=30. URL. https://www.sim-one.ca/community/tip/analyzing-likert-scale-data-rule-n30.

Xie, H., Wang, F., Hao, Y., Chen, J., An, J., Wang, Y., & Liu, H. (2017). The more total cognitive load is reduced by cues, the better retention and transfer of multimedia learning: A meta-analysis and two meta-regression analyses. PloS one, 12(8), e0183884.

Virtual Reality in Special Education: An Application Review

Yi (Joy) Li
joy.li@kennesaw.edu

Zhigang Li
zli8@kennesaw.edu

Chi Zhang
chizhang@kennesaw.edu

College of Computing and Software Engineering
Kennesaw State University
Marietta, GA 30060, USA

Abstract

To investigate the state-of-the-art of virtual reality in special education, we reviewed the related research over the past ten years. Strategies and approaches of the study design have been characterized and categorized based on their research focuses. Both perspectives from the special educators and the students with special needs are addressed. This study reveals that immersive virtual reality is effective in special education, while challenges still remain in this area. We provide insights for future studies and also call for more collaboration among researchers, practitioners, and educators.

Keywords: virtual reality, special education, human-computer interaction.

1. INTRODUCTION

In 2019–2020, 7.3 million students ages 3-21 received special education services, or 14 percent of all public-school students (National Center for Education Statistics, 2022). There were 463,200 special education teachers in the same year to accommodate the students with special needs (U.S. Bureau of Labor Statistics, 2022). In 2019, 44 states reported the special education teacher shortage to the federal government. In the 2022 school year, this number increased to 48 (Gaines, 2022). There is a growing need for qualified special education teachers and effective teaching methods and tools to assist students with different learning disabilities. As the current shortage of special education teachers takes time to get solved, we may explore how technologies fill in and help mitigate the problem.

As a part of immersive technologies and the foundation of the modern "metaverse," virtual reality (VR) technology has seen rapid evolution in the past decade. It is increasingly integrated into various areas for research and applications. Pedagogical theories naturally support emerging interactive media, including immersive technologies. As a result, many educators have explored integrating VR, augmented reality (AR), or mixed reality (MR) into teaching. Four identified attributes made VR so promising in the educational field: customizable context suitable for situated learning (Chiou, 2020); embodied interaction for the presence and control (Johnson-Glenberg, 2018); immersive and interactive scenarios for engagement (Allcoat & Mühlenen, 2018); and affordability in both cost and space (Elliott, Peiris, & Parnin, 2015).

Immersive technologies have been increasingly explored and adopted by mainstream education in curriculums, including K-12 and higher education (Li, Zhang, & Luo, 2021). According to their review, the number of studies on VR is constantly increasing. These technologies are proven capable of stimulating desired outcomes, such as improved attitude, motivation, engagement, learning performance, higher-order thinking (critical thinking), communication, collaboration, and learning experience. However, because of the multiple challenges students with special needs usually face, specialized learning strategies are needed for designing practical applications to help them increase self-efficacy and reach their potential (Buzio, Chiesa, & Toppan, 2017). Despite the growing needs and benefits of using immersive technologies in special education, the number of research reports is still incomparable to that of mainstream education (Carreon, Smith, Mosher, Rao, & Rowland, 2022). This motivated us to conduct a literature review of the state-of-the-art immersive technologies being used in special education.

In search of the reviews on using immersive technologies in special education and inclusive education, we found two recently published literature reviews on AR in educational inclusion and special education (Baragash, Al-Samarraie, Alzahrani, & Alfarraj, 2020; Quintero, Baldiris, Rubira, Cerón, & Velez, 2019). The two reviews state that mobile AR solutions were more adopted due to lower prices and less space needed. In light of the findings, we opted to anchor our review on the advantages and challenges of using VR in special education.

In this paper, we aim to: (1) review and summarize the studies that use virtual reality in assisting in special education, including both teacher training and students with learning disabilities, (2) discuss the current status and challenges of the VR-based approach applied in special education, and (3) provide insight into the education community and provide future directions that will benefit researchers, educators, and students with special needs.

2. BACKGROUNDS

For the best practices in learning and teaching, researchers have defined different learning styles for students that react to stimulus variation in teaching – environmental, emotional, sociological, physiological, and psychological fields (Searson & Dunn, 2001). This led to the need to develop different teaching strategies for the different learning styles (Beck, 2001). In addition, factors such as personality and social factors (Busato, Prins, Elshout, & Hamaker, 2000; Klašnja-Milićević, Vesin, Ivanović, & Budimac, 2011; Mills, 1993), genders (D. Garland & Martin, 2005), intellectual abilities, and mental and physical conditions (Whitely, 1924), all affect the learning process and require different teaching strategies (Chou & Wang, 2000; Mondal, 2013).

When it comes to teaching students with special needs, because of various types of disabilities, students may have one or combined challenges in learning, such as attention, language, spatial abilities, memory, higher reasoning, and knowledge acquisition (Dragoo & Lomax, 2020; Jeffs, 2009a). Therefore, more diverse and customizable teaching strategies to provide a realistic environment are needed while guaranteeing student safety. Prior research has pointed out that VR has the potential to address different disabilities when implementing effective teaching in the special education field (Jeffs, 2009a; Lányi, Geiszt, Károlyi, Tilinger, & Magyar, 2006; Powers & Melissa, 1994). A recent similar review (Carreon et al., 2022) analyzed studies until 2019 and focused on the intervention strategies for K-12 students with disabilities. The findings can be applied to the current research we focus on.

3. LITERATURE REVIEW METHODS

Search strategy

This review explores the approaches and studies in the past decade that integrated VR in teaching and educational training for special education. We discuss the challenges and directions facing using VR in special education for educational research. The period of 2010-2022 was determined because of the current trending head-mounted VR devices and the success of the Oculus Rift Development Kit 1 (Dybsky, 2017). Electronic sources were searched for publications between 2010-2022 through Google Scholar, IEEE Xplore, ACM Digital Library, Scopus (Sage, Springer, Science Direct), PubMed, Taylor & Francis, Wiley, Emerald, and Web of Science. The chosen databases are commonly used in education topics and educational technology. The key search term was ("virtual reality" OR "virtual classroom" OR "virtual environment") AND ("special education" OR "special needs"). The references in the included articles were also screened for additional qualified studies.

The authors independently identified articles

using the same criteria described in the next section and performed cross-checks based on the inclusion criteria.

Eligibility criteria

We used the following inclusion criteria for accumulating the articles:

- Available in full text
- Written in English
- Related to immersive VR in Special Education
- Peer-reviewed conference and journal
- Published after 2010

The following exclusion criteria are used to screeen and filter the articles:

- No books/chapters
- Non-immersive VR
- No applications or studies in diagnosis, therapy, and treatment

4. RESULTS

In this section, we summarize the results of the findings and then elaborate on them in detail by categories.

Study Selection Results

The search yielded 207 potentially relevant articles, and 59 additional articles were identified and screened in addition to the original research. Figure 1 shows the PRISMA flow chart (Liberati et al., 2009) of the search process and the numbers of inclusion and exclusion.

After cross-checking and comparing different special needs categories from various resources (Buzio et al., 2017; Jeffs, 2009a), we adopted the following four categories in our review: individuals with physical disabilities, developmental disabilities, behavioral-emotional disabilities, and sensory impairments. The taxonomy of the identified studies is summarized in Table 1. A more detailed explanation and findings are described in the following sections.

Teacher Training

Even though VR has been trending in education, and many studies have reported positive effects in assisting learning for children with special needs, few studies address teacher training for special education. Among all the articles we have reviewed, VR is used explicitly in teacher training for special education in only three of them (Fraser et al., 2020; K. V. Garland et al., 2012; Ip et al., 2020). The individual summaries are compiled in Table 2.

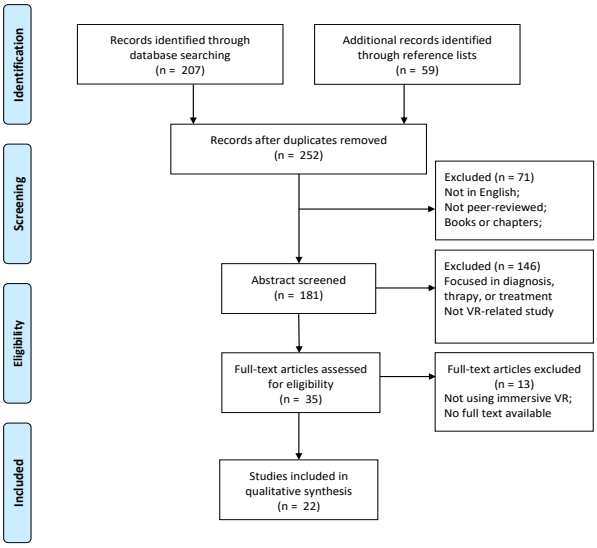


Figure 1: The Prisma Search Process

Garland et al. (K. V. Garland et al., 2012) made

Categories		Reference
Teacher Training		(Fraser et al., 2020; K. V. Garland, Vasquez, & Pearl, 2012; Ip, Li, & Ma, 2020)
Students with conditions	Physical disabilities	(Demers, Martinie, Winstein, & Robert, 2020; Kang & Kang, 2019; Sobota, Korecko, Pastornicky, & Jacho, 2016)
	Developmental disabilities	(Bozgeyikli, Raij, Katkoori, & Alqasemi, 2018; Bradley & Newbutt, 2018; Dechsling et al., 2021; Lorenzo, Newbutt, & Lorenzo-Lledó, 2021; Loup-Escande, Christmann, Damiano, Hernoux, & Richir, 2014; Michalski, Ellison, Szpak, & Loetscher, 2021; Parsons & Cobb, 2011; Tatale, Bhinid, Parmar, & Pcnvar, 2019)
	Behavioral-emotional disabilities	(Bashiri, Ghazisaeedi, & Shahmoradi, 2017; Héctor Cardona-Reyes, Muñoz-Arteaga, Villalba-Condori, & Barba-González, 2021; Hector Cardona-Reyes, Ortiz-Aguinaga, Barba-Gonzalez, & Munoz-Arteaga, 2021; Romero-Ayuso et al., 2021; Seo, Kim, Mundy, Heo, & Kim, 2019)
	Sensory Impairments	(Hrishikesh & Nair, 2016; Yiannoutsou, Johnson, & Price, 2021; Zirzow, 2015)

Table 1: Taxonomy of studies on using immersive VR in special education

use of a popular mixed-reality tool in 2012, TLE TeachLivE, to train teachers to implement a well-justified but rather complicated-to-use method called Discrete Trial Teaching (DTT) for teaching children with autistic spectrum disorders (ASD). The four graduate students who assisted this study demonstrated that using VR can better help coach the teachers to learn practical teaching skills with less negative feelings - tiring, confusing, or frustrating. Eight years later, Fraser et al. (Fraser et al., 2020) decided to repeat the experiment with five special educators who scored lower than average in implementing the DTT. The cohort study of the maintenance probe has shown that the fidelity of implementation has been maintained for at least eight weeks.

Another special educator training approach implemented in Hong Kong involved 76 teachers from 22 local schools and benefited 171 students with ASD (Ip et al., 2020)]. The authors implemented a teacher training program that provided a mechanism to sustain the need of VR-enabled learning for students with ASD.
We also identified a few other teacher training studies that are not designed to target special educators (Alonso et al., 2021; Cárdenas, Álvarez, Romero, & Manero, 2021; Stavroulia et al., 2019; Yun, Park, & Ryu, 2019) but can be extended to special education. The studies mainly focused on strengthening teacher efficacy for handling commonly seen conflicts or incidents on campus. Again, the VR simulated scenarios were proven to substantially impact participants' emotional states. These studies are summarized in Table 3.

Students with Physical Disabilities
Although it was only in the past decade that the advancement and development of portable and affordable VR technologies enabled massive adoption and applications of VR, researchers have long studied the potential benefits of using VR for the intervention and rehabilitation of people with physical disabilities since the 1990s (McComas, Pivik, & Laflamme, 1998). Studies on VR and physical disabilities ranged from VR attributes as an assessment tool, intervention studies, and upper-extremity performance to visual perceptual skills (Laufer & Weiss, 2011). In this study, we identified three articles specific to our topic of using VR for students with physical disabilities for special education. We found two review papers and one study. The summary of the study paper (Sobota et al., 2016) is in Table 4.

In 2016, Sobota et al. (Sobota et al., 2016) presented a virtual reality laboratory that creates an immersive educational environment for children with disabilities. Besides using an HMD such as the Oculus Rift, they also incorporated a 3D scanner to collect data from real-world settings. They have also established a special room with the walls, floor, and ceiling covered with projection screens and a motion capture system to capture a user's movement data. The combination of different technologies allowed the team to adapt the laboratory environment for children with various disabilities.

Three years later, Kang and Kang (Kang & Kang, 2019) reviewed the applications of virtual reality in physical education. They recognized the need to develop a particular curriculum based on the types of disabilities among students and the safety concerns associated with physical activities. As a result, the Korean government implemented a VR Sports Classroom at ten schools for physical sports education for disabled children.

More recently, in the context of the COVID-19 pandemic, Demers et al. (Demers et al., 2020) examined the use of low-cost virtual reality video games to provide students with physical disabilities an opportunity for motor skills learning in a home environment. The authors reviewed the evidence of VR technologies and their effectiveness in rehabilitating people with disabilities. However, they also pointed out that clinicians play an essential role in adapting and selecting VR platforms and games. There is still a lack of resources and tools to support clinicians in making such decisions about VR adoption and integration.

Students with Developmental Disabilities
Examples of developmental disabilities include autism, Down syndrome, dyslexia, processing disorders, and more. The findings are summarized in the following subsections.

Autistic spectrum disorders. As the main focus and typical representation of developmental disorders, Autistic Spectrum Disorders (ASD) have been extensively researched with innovative methods for intervention and training. We have gathered 44 articles and five review papers related to using VR for educational purposes for ASD. Instead of going through each article, we decided to focus this section on getting insights by reviewing the reviews. The research focus of the studies is summarized in Figure 2.

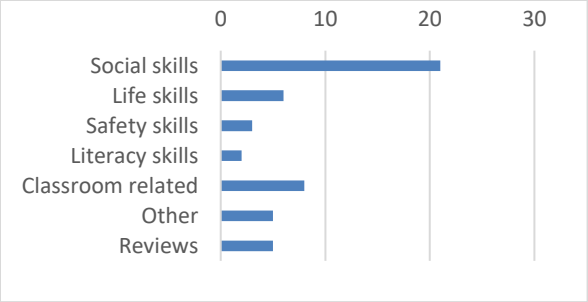


Figure 2: Number of articles by research focus

Although considered a childhood condition, ASD impairments are generally life-long, and 1 in 44 children have autism in the US (CDC, 2022a). People with ASD typically suffer from deficits in communication, emotional capacity, social interactions, and repetitive behavior patterns ("Autism Spectrum Disorder," n.d.; CDC, 2022b).

In our categorization, articles related to social skills include emotional development, communication skills, and attention skills; life skills include travel training, shopping, adaptive skills, and interviewing skills; safety skills include street crossing and fire safety; literacy skills look at word processing and conceptual comprehension. Classroom-related research focuses on virtual classroom settings for learning, including student engagement and anxiety or disruptive behavior management.

Parsons and Cobb (Parsons & Cobb, 2011) reviewed studies ten years prior to 2010 that discussed using virtual environments in special education for children with ASD, particularly in improving social skills. Since immersive VR devices (especially HMD) have not been made so portable and convenient at the time, evidence is limited to support the full potential of using VR for ASD in teaching.

Bozgeyikli et al. (Bozgeyikli et al., 2018) summarized studies using VR (immersive and non-immersive) prior to 2018 that targeted ASD teaching or educational training. The research targets three types of skills: social, life and safety skills. The mainstream use of immersive VR includes HMD and CAVE. Ten studies using immersive VR, with three to four under each skill category, yielded positive training results and successful transfer of learned skills to real life. Design principles were proven effective for autism and aligned with these for typical inclusive education, such as embodied learning, feeling of control, goal orientation, repetition, task complexity, and rewards with feedback. On the other hand, one common limitation of the discussed studies is insufficient fidelity due to the small average number of participants (only 7).

Bradley and Newbutt (Bradley & Newbutt, 2018) raised questions about the robustness of VR applications designed for ASD children despite the typical optimism in treatment. Thus, they identified and reviewed six studies that used HMD-VR with empirical study data. Similar to the above review work, the lack of participants hinders the positive conclusions. Furthermore, four out of the six studies did not include a control group, limiting the extent of the confidence in the findings. The most negative effects being reported are fatigue and cybersickness. The authors pointed out that because of the heightened sensory concerns of the autistic population, extra care and an ethical framework should be established for designing suitable VR for children with ASD.

Dechsling et al. (Dechsling et al., 2021) reviewed VR and AR studies specifically for improving social skills for people with ASD. Forty-nine studies were identified, aligning with our observation that social skills are the most targeted in using VR to help autism. These studies show that young children, adults, and female participants are less targeted. Rigorous research designs and evidence-based study strategies are needed.

Lorenzo et al. (Lorenzo et al., 2021) analyzed the global trends in using VR for people with ASD. The results show that the world growth of keywords has aligned with the use of virtual reality in ASD has rapidly increased since 2011. Yet, special education has not been addressed as much as other normal K-12 or higher education.

Other developmental disabilities. The use of VR to help people with other developmental disabilities is addressed very little compared to the research interest in ASD in the literature. We found two review papers and three studies; the papers (de Vasconcelos et.al., 2020, 2020; Loup-Escande et al., 2014; Tremblay et. al., 2014) are summarized in Table 4.

Loup-Escande et al. (Loup-Escande et al., 2014) designed a VR learning software with three tasks of dishwashing activities for students with congenital mental disabilities to determine whether using touchscreen or mouse interactions is better for the targeted audience in learning life skills. According to both performance and post-session interviews, the touchscreen is preferred for similar study or application designs.

Michalski et al. (Michalski et al., 2021) reviewed

eight studies using virtual interactive training agents for vocational skills in people with neurodevelopmental disorders. The studies proved that participants could transfer vocational skills from the experimental session to real-world settings.

As for targeting students with dyslexia, Tatale et al. (Tatale et al., 2019) reviewed 11 articles on using VR to educate students with learning disabilities. Based on a selection of four of the studies, the authors proposed a system that uses WebVR for text visualization to teach and test students. The tests provided evidence for customizable courses accordingly. Unfortunately, this is just a system outline without empirical study for assessment.

Other studies that targeted learning disabilities (LD) with visual-motor skills (Tremblay et al., 2014) and literacy skills (de Vasconcelos et al., 2020) both reported positive evaluations for the focus group. However, the adult participants with LD in (Tremblay et al., 2014) did not improve learning as much as the control group without LD but also used VR, which may have provided a differentiation method.

Students with Behavioral-emotional Disabilities

Behavioral or emotional disabilities is an umbrella term that includes a wide range of conditions, including but not limited to attention deficit hyperactivity disorder (ADHD), bipolar, and oppositional defiance disorder. Among these conditions, ADHD grabs the most attention from researchers. In this study, we selected five articles specific to our topic of using VR with students with behavioral or emotional disabilities. Among them, two are literature reviews and meta-analyses, and three are studies (Héctor Cardona-Reyes et al., 2021; Hector Cardona-Reyes et al., 2021; Seo et al., 2019). The study papers are summarized in Table 4.

In 2017, Bashiri et al. (Bashiri et al., 2017) conducted a literature review from 2000 to 2017 on the use of VR for the rehabilitation of children with ADHD. In their review, the authors gathered 12 research studies on the application of VR classroom technology. The topics of these studies ranged from cybersickness to student performance and attention. With the growing research and evidence in VR applications for children with special needs, the authors highlighted the opportunities VR systems present in meeting the educational needs of children with ADHD.

Seo et al. (Seo et al., 2019) published a preliminary study on an evaluation of student attention in social situations in a joint attention virtual reality classroom. The research team conducted an initial experiment with healthy individuals and found that the joint attention VR classroom promoted attentional processes through virtual social interaction. The authors concluded that the joint attention VR classroom can be an important tool to facilitate social interactions for school-aged children with ADHD.

Cardona-Reyes et al. presented a user-centered Lean UX process model for designing a VR environment for ADHD children (Héctor Cardona-Reyes et al., 2021; Hector Cardona-Reyes et al., 2021). A total of 25 elementary school children participated in the study using the developed Attention VR virtual reality environment. Among them, seven children are with ADHD, 1 with Asperger's syndrome, and the rest 17 are regular children. Their results revealed a positive attitude towards the VR environment from the students. It can be a viable option to extend educational opportunities to children with ADHD with the limitations presented by the pandemic.

In 2021, Romero-Ayuso et al. (Romero-Ayuso et al., 2021) conducted a meta-analysis of research studies on the effectiveness of VR-based intervention for children with ADHD. After rounds of exclusion, they selected four articles for the meta-analysis. The authors found that most of the VR studies regarding ADHD focused on the validation of the assessment of attention instead of cognitive rehabilitation. Through their analysis, the authors concluded that "VR-based interventions help to improve the cognitive performance of children and adolescents with ADHD in vigilance and sustained-attention tasks, reducing the number of omissions, and increasing the number of correct responses to the target stimuli with large effect size" (Romero-Ayuso et al., 2021).

Students with Sensory Impairment

Sensory impairment refers to individuals with challenges in "one or more of the three senses-vision, touch, and hearing" (Jeffs, 2009b). Similar to other categories of VR research in this study, evidence of using VR to help sensory impaired students appeared in the early 2000s (Passig & Eden, 2003). Researchers agree that VR technologies help break down barriers for students with hearing impairment and allow them to practice essential skills needed in the real-world (Jeffs, 2009b; Lányi et al., 2006; Passig & Eden, 2003). In this study, we identified three articles specific to our topic of using VR for

students with sensory impairment for special education. Among the three, one is a review article, and two are studies (Hrishikesh & Nair, 2016; Yiannoutsou et al., 2021). The study papers are summarized in Table 4.

In an article published by Zirzow (Zirzow, 2015), the author discussed the use of signing avatars to support students with hearing loss. A signing avatar refers to "an animated 3D model of a virtual human that presents messages in sign language" (De Martino et al., 2017). It provides an alternative to spoken language and offers an opportunity to break down the learning barrier for students with hearing impairment. Zirzow (Zirzow, 2015) examined current applications of signing avatars and emerging technologies such as SMILE (Science and Math in an Immersive Environment), which offers an immersive virtual learning environment for elementary-level school children. The author noted that educators have to provide explicit instructions to students with disabilities for the technology to be used appropriately. She also recognized the need for more research on non-manual signals such as body motion and facial expression of the signed avatars to realize their potential fully.

Hrishikesh and Nair (Hrishikesh & Nair, 2016) developed a virtual interactive learning system that provides students, particularly the hearing impaired and vocally challenged, with an immersive learning experience. The system uses a combination of VR and AR along with Microsoft Kinect for motion tracking. A prototype system was made and tested in local schools. Two groups of students were chosen, with one group taught using the virtual interactive learning system and the other without. The students were given a pop quiz and survey to compare the learning curve and learning experience. Their results revealed that students taught using the virtual learning system had higher accuracy in answering the questions in the quizzes. Students also experienced a deeper immersion and engagement with the subject.

Yiannoutsou et al. (Yiannoutsou et al., 2021) realized the need for more research on using VR technologies with visually impaired students. They created a system using HTC Vive to provide visually impaired students with a VR environment. The speakers from the system and the vibration from the controllers provided students with both audio and tactile feedback while completing tasks. Seven visually impaired children were recruited and participated in the experiment. In their conclusion, the authors noted that VR technologies disrupt existing school practices. However, the need for dedicated physical space for the VR environment presents a challenge for schools to adopt these systems due to spatial constraints.

5. DISCUSSION

Teacher training using VR is still under-researched, considering the fact that more and more students have access to VR applications, and researchers have been continuing to put effort into using VR in education. This trend may lead to children being more familiar with VR than their teachers (Ainge, 1997), while teachers find it challenging to implement justified teaching methods dedicated to special needs with VR tools. Therefore, more VR-based teacher training is needed. Subsection 4.2 on teacher training indicated two directions for future study designs. The first direction is to develop VR scenarios to train teachers with justified teaching methods or skills for better serving children with special needs, such as empathy, communication, classroom management, and more. The other direction is to prepare teachers or pre-service teachers with the knowledge of using VR and their peripheral accessories and encourage them to integrate existing VR content into the curriculum. Furthermore, collaborating with developers to customize suitable VR applications for their students would also help students in need.

As noted in the selected articles reviewed in this paper, VR, as an emerging technology, enables an unprecedented glimpse inside special education and fascinates researchers, educators, and practitioners with its broad potential to support the educational needs of children with disabilities. Among different types of disabilities, ASD has attracted the most attention from researchers with a leading number of research publications. Trailing behind are studies for children with ADHD and other types of disabilities.

As demonstrated in one of the meta-analysis studies (Romero-Ayuso et al., 2021), a large majority of studies on VR for special education and VR for people with disabilities are non-experimental. Despite the general positive attitude and recommendations for adopting VR for people with special needs, there is still a lack of experimental studies with control groups to provide more affirmative conclusions. Furthermore, interdisciplinary collaborations involving practitioners in the field (such as psychiatrists or consultants), along with professionals in technology development and special educators directly working with students

with conditions should be encouraged with more resources and opportunities.

6. CONCLUSION AND FUTURE WORK

In this paper, we reviewed the prior studies on how effective virtual reality is applied in special education.

Regarding aims 1 and 2 mentioned in Introduction, after synthesizing the trends and challenges, we identified two audience perspectives as, teachers' and students', and four categories of special needs - physical disabilities, developmental disabilities, behavioral-emotional disabilities, and sensory impairments. We collected individual studies for each category and investigated the recent studies and findings. The studies were examined from teachers' and students' perspectives. We also discussed how effective VR is used for improving various skill deficits.

Regarding aim 3, with this review of VR for special education, we hope to encourage more development of immersive technologies, especially VR-based, based on the existing and proven theories in special education. We urge more attention and support to be provided to encourage interdisciplinary collaborations and more participants in future studies.

7. REFERENCES

Ainge, D. (1997). Virtual Reality in Schools: The Need for Teacher Training. *Innovations in Education and Training International*, *34*(2), 114–118. doi: 10.1080/1355800970340206

Allcoat, D., & Mühlenen, A. von. (2018). Learning in virtual reality: Effects on performance, emotion and engagement. *Research in Learning Technology*, *26*. doi: 10.25304/rlt.v26.2140

Alonso, S., López, D., Puente, A., Romero, A., Álvarez, I. M., & Manero, B. (2021). *Evaluation of a Motion Capture and Virtual Reality Classroom for Secondary School Teacher Training*. 6.

Autism Spectrum Disorder. (n.d.). Retrieved May 23, 2022, from National Institute of Mental Health (NIMH) website: https://www.nimh.nih.gov/health/topics/autism-spectrum-disorders-asd

Baragash, R. S., Al-Samarraie, H., Alzahrani, A. I., & Alfarraj, O. (2020). Augmented reality in special education: A meta-analysis of single-subject design studies. *European Journal of Special Needs Education*, *35*(3), 382–397. doi: 10.1080/08856257.2019.1703548

Bashiri, A., Ghazisaeedi, M., & Shahmoradi, L. (2017). The opportunities of virtual reality in the rehabilitation of children with attention deficit hyperactivity disorder: A literature review. *Korean Journal of Pediatrics*, *60*(11), 337–343. doi: 10.3345/kjp.2017.60.11.337

Beck, C. R. (2001). Matching teaching strategies to learning style preferences. *The Teacher Educator*, *37*(1), 1.

Bozgeyikli, L., Raij, A., Katkoori, S., & Alqasemi, R. (2018). A Survey on Virtual Reality for Individuals with Autism Spectrum Disorder: Design Considerations. *IEEE Transactions on Learning Technologies*, *11*(2), 133–151. doi: 10.1109/TLT.2017.2739747

Bradley, R., & Newbutt, N. (2018). Autism and virtual reality head-mounted displays: A state of the art systematic review. *Journal of Enabling Technologies*, *12*(3), 101–113. doi: 10.1108/JET-01-2018-0004

Busato, V. V., Prins, F. J., Elshout, J. J., & Hamaker, C. (2000). Intellectual ability, learning style, personality, achievement motivation and academic success of psychology students in higher education. *Personality and Individual Differences*, *29*(6), 1057–1068. doi: 10.1016/S0191-8869(99)00253-6

Buzio, A., Chiesa, M., & Toppan, R. (2017). Virtual Reality for Special Educational Needs. *Proceedings of the 2017 ACM Workshop on Intelligent Interfaces for Ubiquitous and Smart Learning*, 7–10. New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3038535.3038541

Cárdenas, M. M., Álvarez, I. M., Romero, A., & Manero, B. (2021). A Teacher Training Proposal for Classroom Conflict Management through Virtual Reality. *2021 International Conference on Advanced Learning Technologies (ICALT)*, 373–375. doi: 10.1109/ICALT52272.2021.00120

Cardona-Reyes, Héctor, Muñoz-Arteaga, J., Villalba-Condori, K., & Barba-González, M. L. (2021). A Lean UX Process Model for Virtual Reality Environments Considering ADHD in Pupils at Elementary School in COVID-19 Contingency. *Sensors*, *21*(11), 3787. doi: 10.3390/s21113787

Cardona-Reyes, Hector, Ortiz-Aguinaga, G., Barba-Gonzalez, M. L., & Munoz-Arteaga, J.

(2021). User-Centered Virtual Reality Environments to Support the Educational Needs of Children With ADHD in the COVID-19 Pandemic. *IEEE Revista Iberoamericana de Tecnologias Del Aprendizaje*, *16*(4), 400–409. doi: 10.1109/RITA.2021.3135194

Carreon, A., Smith, S. J., Mosher, M., Rao, K., & Rowland, A. (2022). A Review of Virtual Reality Intervention Research for Students With Disabilities in K–12 Settings. *Journal of Special Education Technology*, *37*(1), 82–99. doi: 10.1177/0162643420962011

CDC. (2022a, March 2). Data and Statistics on Autism Spectrum Disorder | CDC. Retrieved May 23, 2022, from Centers for Disease Control and Prevention website: https://www.cdc.gov/ncbddd/autism/data.html

CDC. (2022b, March 28). Signs & Symptoms | Autism Spectrum Disorder (ASD) | NCBDDD | CDC. Retrieved May 23, 2022, from Centers for Disease Control and Prevention website: https://www.cdc.gov/ncbddd/autism/signs.html

Chiou, H.-H. (2020). The impact of situated learning activities on technology university students' learning outcome. *Education + Training*, *63*(3), 440–452. doi: 10.1108/ET-04-2018-0092

Chou, H.-W., & Wang, T.-B. (2000). The influence of learning style and training method on self-efficacy and learning performance in WWW homepage design training. *International Journal of Information Management*, *20*(6), 455–472. doi: 10.1016/S0268-4012(00)00040-2

De Martino, J. M., Silva, I. R., Bolognini, C. Z., Costa, P. D. P., Kumada, K. M. O., Coradine, L. C., … De Conti, D. F. (2017). Signing avatars: Making education more inclusive. *Universal Access in the Information Society*, *16*(3), 793–808. doi: 10.1007/s10209-016-0504-x

de Vasconcelos, D. F. P., Júnior, E. A. L., de Oliveira Malaquias, F. F., Oliveira, L. A., & Cardoso, A. (2020). A Virtual Reality based serious game to aid in the literacy of students with intellectual disability: Design principles and evaluation. *Technology and Disability*, *32*(3), 149–157.

Dechsling, A., Orm, S., Kalandadze, T., Sütterlin, S., Øien, R. A., Shic, F., & Nordahl-Hansen, A. (2021). Virtual and Augmented Reality in Social Skills Interventions for Individuals with Autism Spectrum Disorder: A Scoping Review. *Journal of Autism and Developmental Disorders*. doi: 10.1007/s10803-021-05338-5

Demers, M., Martinie, O., Winstein, C., & Robert, M. T. (2020). Active Video Games and Low-Cost Virtual Reality: An Ideal Therapeutic Modality for Children With Physical Disabilities During a Global Pandemic. *Frontiers in Neurology*, *11*. Retrieved from https://www.frontiersin.org/article/10.3389/fneur.2020.601898

Dragoo, K. E., & Lomax, E. (2020). The Individuals with Disabilities Education Act: A Comparison of State Eligibility Criteria. CRS Report R46566, Version 5. In *Congressional Research Service*. Congressional Research Service. Retrieved from https://eric.ed.gov/?id=ED610722

Dybsky, D. (2017, March 1). The History of Virtual Reality: Ultimate Guide. Part 2 | TESLASUIT blog. Retrieved September 26, 2020, from TESLASUIT website: https://teslasuit.io/blog/history-virtual-reality-ultimate-guide-part-2/

Elliott, A., Peiris, B., & Parnin, C. (2015). Virtual Reality in Software Engineering: Affordances, Applications, and Challenges. *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*, *2*, 547–550. doi: 10.1109/ICSE.2015.191

Fraser, D. W., Marder, T. J., deBettencourt, L. U., Myers, L. A., Kalymon, K. M., & Harrell, R. M. (2020). Using a Mixed-Reality Environment to Train Special Educators Working With Students With Autism Spectrum Disorder to Implement Discrete Trial Teaching. *Focus on Autism and Other Developmental Disabilities*, *35*(1), 3–14. doi: 10.1177/1088357619844696

Gaines, L. V. (2022, April 20). Students with disabilities have a right to qualified teachers—But there's a shortage. *NPR*. Retrieved from https://www.npr.org/2022/04/20/1092337446/special-education-teacher-shortage

Garland, D., & Martin, B. (2005). Do Gender and Learning Style Play a Role in How Online Courses Should Be Designed? *Journal of Interactive Online Learning*, *4*.

Garland, K. V., Vasquez, E., & Pearl, C. (2012). Efficacy of Individualized Clinical Coaching in a Virtual Reality Classroom for Increasing Teachers' Fidelity of Implementation of Discrete Trial Teaching. *Education and*

Training in Autism and Developmental Disabilities, *47*(4), 502–515.

Hrishikesh, N., & Nair, J. J. (2016). Interactive learning system for the hearing impaired and the vocally challenged. *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 1079–1083. doi: 10.1109/ICACCI.2016.7732188

Ip, H., Li, C., & Ma, P. K. (2020). A Teacher Training Approach to Sustain Virtual Reality Enabled Learning in The Inclusive Education Setting for Children With Autism Spectrum Disorder. *INTED2020: 14th International Technology, Education and Development Conference*, 582–588. IATED Academy. doi: 10.21125/inted.2020.0230

Jeffs, T. L. (2009a). Virtual Reality and Special Needs. *Themes in Science and Technology Education*, *2*, 253–268.

Jeffs, T. L. (2009b). Virtual Reality and Special Needs. *Themes in Science and Technology Education*, (Special Issue), 253–268.

Johnson-Glenberg, M. C. (2018). Immersive VR and Education: Embodied Design Principles That Include Gesture and Hand Controls. *Frontiers in Robotics and AI*, *5*, 81. doi: 10.3389/frobt.2018.00081

Kang, S., & Kang, S. (2019). The study on the application of virtual reality in adapted physical education. *Cluster Computing*, *22*(S1), 2351–2355. doi: 10.1007/s10586-018-2254-4

Klašnja-Milićević, A., Vesin, B., Ivanović, M., & Budimac, Z. (2011). E-Learning personalization based on hybrid recommendation strategy and learning style identification. *Computers & Education*, *56*(3), 885–899. doi: 10.1016/j.compedu.2010.11.001

Lányi, C. S., Geiszt, Z., Károlyi, P., Tilinger, Á., & Magyar, V. (2006). Virtual Reality in special needs early education. *International Journal of Virtual Reality*, *5*(3), 1–10.

Laufer, Y., & Weiss, P. (Tamar) L. (2011). Virtual Reality in the Assessment and Treatment of Children With Motor Impairment: A Systematic Review. *Journal of Physical Therapy Education*, *25*(1), 59–71. doi: http://dx.doi.org/10.1097/00001416-201110000-00011

Li, Y. (Joy), Zhang, C., & Luo, H. (Irene). (2021). Using Mixed Reality in K-12 Education: A Literature Review. *AMCIS 2021 Proceedings*. Retrieved from https://aisel.aisnet.org/amcis2021/sig_hci/sig_hci/22

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., … Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *Journal of Clinical Epidemiology*, *62*(10), e1–e34. doi: 10.1016/j.jclinepi.2009.06.006

Lorenzo, G., Newbutt, N., & Lorenzo-Lledó, A. (2021). Global trends in the application of virtual reality for people with autism spectrum disorders: Conceptual, intellectual and the social structure of scientific production. *… of Computers in Education*. doi: 10.1007/s40692-021-00202-y

Loup-Escande, E., Christmann, O., Damiano, R., Hernoux, F., & Richir, S. (2014). Virtual reality learning software for individuals with intellectual disabilities: Comparison between touchscreen and mouse interactions. *International Journal of Child Health and Human Development*, *7*(4), 415–424.

McComas, J., Pivik, J., & Laflamme, M. (1998). Children's Transfer of Spatial Learning from Virtual Reality to Real Environments. *CyberPsychology & Behavior*, *1*(2), 121–128. doi: 10.1089/cpb.1998.1.121

Michalski, S. C., Ellison, C., Szpak, A., & Loetscher, T. (2021). Vocational Training in Virtual Environments for People With Neurodevelopmental Disorders: A Systematic Review. *Frontiers in Psychology*, *12*, 627301. doi: 10.3389/fpsyg.2021.627301

Mills, C. J. (1993). Personality, Learning Style and Cognitve Style Profiles of Mathematically Talented Students. *European Journal of High Ability*, *4*(1), 70–85. doi: 10.1080/0937445930040108

Mondal, P. (2013, August 22). 7 Important Factors that May Affect the Learning Process. Retrieved May 9, 2022, from Your Article Library website: https://www.yourarticlelibrary.com/learning/7-important-factors-that-may-affect-the-learning-process/6064

National Center for Education Statistics. (2022). *Students With Disabilities*. U.S. Department of Education, Institute of Education Sciences. Retrieved from U.S. Department of Education, Institute of Education Sciences

website: https://nces.ed.gov/programs/coe/indicator/cgg

Parsons, S., & Cobb, S. (2011). State-of-the-art of virtual reality technologies for children on the autism spectrum. *European Journal of Special Needs Education*, *26*(3), 355–366. doi: 10.1080/08856257.2011.593831

Passig, D., & Eden, S. (2003). Cognitive intervention through virtual environments among deaf and hard-of-hearing children. *European Journal of Special Needs Education*, *18*(2), 173–182. doi: 10.1080/0885625032000078961

Powers, D. A., & Melissa, D. (1994). Special Education and Virtual Reality. *Journal of Research on Computing in Education*, *27*(1), 111–121. doi: 10.1080/08886504.1994.10782120

Quintero, J., Baldiris, S., Rubira, R., Cerón, J., & Velez, G. (2019). Augmented Reality in Educational Inclusion. A Systematic Review on the Last Decade. *Frontiers in Psychology*, *10*, 1835. doi: 10.3389/fpsyg.2019.01835

Romero-Ayuso, D., Toledano-González, A., Rodríguez-Martínez, M. del C., Arroyo-Castillo, P., Triviño-Juárez, J. M., González, P., … Segura-Fragoso, A. (2021). Effectiveness of Virtual Reality-Based Interventions for Children and Adolescents with ADHD: A Systematic Review and Meta-Analysis. *Children*, *8*(2), 70. doi: 10.3390/children8020070

Searson, R., & Dunn, R. (2001). The learning-style teaching model. *Science and Children*, *38*(5), 22–26.

Seo, S., Kim, E., Mundy, P., Heo, J., & Kim, K. K. (2019). Joint attention virtual classroom: A preliminary study. ncbi.nlm.nih.gov. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/pmc6504768/

Sobota, B., Korecko, S., Pastornicky, P., & Jacho, L. (2016). Virtual-reality technologies in the process of handicapped school children education. *2016 International Conference on Emerging ELearning Technologies and Applications (ICETA)*, 321–326. Starý Smokovec, High Tatras, Slovakia: IEEE. doi: 10.1109/ICETA.2016.7802077

Stavroulia, K. E., Christofi, M., Baka, E., Michael-Grigoriou, D., Magnenat-Thalmann, N., & Lanitis, A. (2019). Assessing the emotional impact of virtual reality-based teacher training. *The International Journal of Information and Learning Technology*, *36*(3), 192–217. doi: 10.1108/IJILT-11-2018-0127

Tatale, S., Bhinid, N., Parmar, R., & Pcnvar, S. (2019). A review on Virtual Reality for educating students with learning disabilities. *2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, 1–6. Pune, India: IEEE. doi: 10.1109/ICCUBEA47591.2019.9128570

Tremblay, L., Chebbi, B., Bouchard, S., Cimon-Lambert, K., & ... (2014). Learning disabilities and visual-motor skills; comparing assessment from a hapto-virtual reality tool and Bender-Gestalt test. *Virtual Reality*. doi: 10.1007/s10055-014-0242-4

U.S. Bureau of Labor Statistics. (2022). *Special Education Teachers*. Retrieved from https://www.bls.gov/ooh/education-training-and-library/special-education-teachers.htm

Whitely, P. L. (1924). The Dependence of Learning and Recall upon Prior Mental and Physical Conditions. *Journal of Experimental Psychology*, *7*(6), 420–428. doi: 10.1037/h0071701

Yiannoutsou, N., Johnson, R., & Price, S. (2021). Non visual Virtual Reality: Considerations for the Pedagogical Design of Embodied Mathematical Experiences for Visually Impaired Children. *Educational Technology & Society*, *24*(2), 151–163.

Yun, H., Park, S., & Ryu, J. (2019, March 18). *Exploring the influences of immersive virtual reality pre-service teacher training simulations on teacher efficacy*. 2112–2116. Association for the Advancement of Computing in Education (AACE). Retrieved from https://www.learntechlib.org/primary/p/207938/

Zirzow, N. K. (2015). Signing Avatars: Using Virtual Reality to Support Students with Hearing Loss. *Rural Special Education Quarterly*, *34*(3), 33–36. doi: 10.1177/875687051503400307

APPENDIX A
Summary of Selected Papers

Ref.	Methods	Teaching skills	Audience	Study design	Participants	Outcomes
(Fraser et al., 2020)	Implemented a justified teaching method in mixed reality to train special educators	Implementing Discrete Trial Teaching (DTT) method.	Special educators of children with ASD	A baseline, two intervention phases comparing didactic training vs. mixed reality training; Videos recorded; maintenance probes for up to 8 weeks.	N=5 lowest scores out of 15 special educators	Participants were able to implement DTT with fidelity in their own classrooms after an hour-long session. They maintained their fidelity of implementation up to 8 weeks after intervention.
(K. V. Garland et al., 2012)	Using TeachLivE to train to implement DTT	Implementing DTT method	Special educators of children with ASD	Multiple-baseline design; coaching with feedback and demonstration served as the independent variable	N=4 graduate students	Coaching with feedback and demonstration lead to functional relationships and fidelity; avatar to target specific skills -> less tiring, confusing, or frustrating.
(Ip et al., 2020)	Hands-on practice with a developed VR learning environment with six modules	CAVE VR and HMD VR practice	Special educators of children with ASD	Three sessions covered 6 VR modules; introduction, practice, and protocols; Divided into groups to practice using VR. Post-questionnaires	N=76 teachers from 22 local schools	Teachers reported continuously using the VR modules to teach students with ASD, who indicated the VR sessions were fun and enjoyable

Table 2: **Teacher training using VR for special educators summary**

Ref.	Methods	Teaching skills	Audience	Study design	Participants	Outcomes
(Alonso et al., 2021)	Manage disruptive situations	ClassroomVR-MotionCapture (CVR-MC)	Secondary school teachers	Analyze users' tone of voice and the substance of their speech, and their gaze and corporal movements	N=14 education professionals	The emotions detected through body expressions did not match the self-reported feeling
(Stavroulia et al., 2019)	Simulated school incidents based on real-life; provides teacher, student drug user and student observer perspectives	School incidents management	Teachers facing students incidents at school	Questionnaires for Empathy scale and mood state scales, Fitbit wristband, EEG	N=25; 9 non-teacher; 6 higher-education; 10 K-12 teachers	VR has a strong impact on participants' emotional states
(Yun et al., 2019)	Virtual classroom with three misbehaved student scenarios	Teacher efficacy	Pre-service teachers	24-item OSTES survey questionnaire; 7-point Likert scale to measure teacher efficacy	N=75 undergrad pre-service teachers	Teacher efficacy was significantly associated with different simulated scenarios
(Cárdenas et al., 2021)	Virtual classroom with students' conflict scenarios	Classroom conflict management	Secondary school teachers	Voice tone, distance from students, and words used were gathered; virtual students' behaviors react to them	N/A	Did not report assessment

Table 3: Teacher training using VR that can be adopted in special education

Ref.	Methods	Teaching skills	Audience	Study design	Participants	Outcomes
(Sobota et al., 2016)	Develop a combination of VR and a smart environment	General	Children with physical disabilities	N/A	N/A	Combining technologies allowed the team to adapt to the lab environment for children with disabilities.
(Loup-Escande et al., 2014)	Comparing tactile touchscreen vs. mouse in "Apticap"	Dish-washing activity	Mental disabilities	Identification, questionnaire, post-session interview	N=6 (2F, 4M) w/ congenital mental deficiency	participants finished tasks faster with the touchscreen than with the mouse
(Seo et al., 2019)	Incorporate social attention among participants and virtual teachers	General	Children with ADHD	ANOVA analysis compares accuracy, commission & omission err, response time & variability, head movements	N=58, 25 for pilot and 33 for main studies	The VR environment promoted attentional processing.
(Hector Cardona-Reyes et al., 2021)	VR learning environment for ADHD students	General	Children with ADHD	Questionnaire and data generated by the system	N=25 children with ADHD	Positive perception and satisfaction from the children.
(Héctor Cardona-Reyes et al., 2021)	A UX design process for VR	General	Children with ADHD	Questionnaire and data generated by the system	N=16 children, 7 with ADHD, 9 control	Positive preliminary results.
(Hrishikesh & Nair, 2016)	Interactive learning system for students with hearing and vocal disabilities	General	Students with hearing loss and vocal challenges	Pop quiz after the learning session.	N=92, randomly selected. Focus group used VR, control group without	Students taught using the VR environment can recall the concept more accurately.
(Yiannoutsou et al., 2021)	Non-visual VR for visually impaired children.	Cartesian coordinates	Visually impaired children	Video recordings, discussions, reflections	N=7	The VR environment helps to support the transfer of knowledge
(Tremblay et al., 2014)	Paper-and-pencil vs. haptic-visual VR motor skill assessment	Visual-motor skills	Students with learning disabilities (LD)	Bender-Gestalt test	N=22 undergrads, 11 with LD, 11 in the control group	Adults with LD did not improve learning as much as the control group using VR; can differentiate the groups
(de Vasconcelos et al., 2020)	VR game validated by special education professionals used in two inclusive schools	Literacy skills	Students with intellectual disability (ID)	Two qualitative questionnaires for students and teachers	N=8 elementary students with ID	Positive evaluation for using VR to teach literacy skills in students with ID

Table 4: Studies using VR targeting students with disabilities