ANALYSES OF THE TCU DRUG SCREEN 5:

USING AN ITEM RESPONSE THEORY MODEL

WITH A SAMPLE OF JUVENILE JUSTICE YOUTH

by

AMANDA LEE WIESE

Bachelor of Science, 2015
California Lutheran University
Thousand Oaks, California

Master of Science, 2018
University of Texas at Dallas
Richardson, Texas

Submitted to the Graduate Faculty of the
College of Science and Engineering
Texas Christian University
in partial fulfillment of the requirements
for the degree of

Master of Science

May 2020

ACKNOWLEDGEMENTS

I wish to express my deepest gratitude to my thesis advisor and mentor, Professor Kevin Knight, Director of the Institute of Behavioral Research, who guided and pushed me to strive for excellence and persevere even when things got tough. His persistent help can be credited for ensuring this project reached its goal.

I would like to recognize the invaluable assistance that my thesis committee provided during this study: Professors Cathy Cox, Danica Kalling Knight, George Joe, and Mary Hargis. Without their support and insights, this project could not have been realized.

The support from the entire staff at the Institute of Behavioral Research is truly appreciated. Their encouragement inspired me to keep reaching for bigger and better things.

I wish to acknowledge the support and boundless love of my family, my mother, Barbara; my father, Christopher; and my brother, Ryan. They kept me going and this work would not have been possible without their belief in me.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ANALYSES OF THE TCU DRUG SCREEN 5:
USING AN ITEM RESPONSE THEORY MODEL
WITH A SAMPLE OF JUVENILE JUSTICE YOUTH

Identifying youth with substance use disorders (SUDs) is the first step in linking them with the treatment services they need. The current study starts with outlining the importance of the need for juvenile justice (JJ) agencies to administer validated, evidence-based screening instruments for SUDs. Next, an overview of the analytic approaches used in the current study is given, including classical test theory (CTT) and item response theory (IRT). The current study contrasts these analytic approaches and compares their effectiveness in identifying and classifying SUDs using an evidence-based screener, the Texas Christian University Drug Screen 5 (TCU DS 5), among a sample of JJ-involved youth.

Section I: Introduction

Substance use (SU) has long been known to play an intimate role in youth involvement within the JJ system. It is estimated that 45% to 65% of JJ-involved individuals in the U.S. meet clinical diagnostic criteria for having a SUD (Dennis et al., 2009). Youth involved in the JJ system are nine times more likely to develop a SUD compared to their adolescent counterparts who do not come into contact with the JJ system (Center for Behavioral Health and Quality, 2016). The prevalence of SU among this population puts them at a heightened risk of mental health issues such as suicidality (Tapia et al., 2016), human immunodeficiency virus (HIV) and other sexually transmitted infections (Donenberg et al., 2015), and criminal recidivism (Henggeler et al., 2002). The JJ system is uniquely positioned to prevent, identify, and treat SUDs among this vulnerable population.

The first step in identifying JJ-involved youth with a SUD is to administer an evidence-based screening instrument, as indicated by the Juvenile Justice Behavioral Health Services Cascade (Belenko et al., 2017). Every individual who enters the JJ system should be screened in a timely manner using a validated screening instrument that provides clinically meaningful results to indicate the severity of SU problems. A comprehensive assessment is then administered to individuals who score above a certain threshold on the screener. The information from the screening and assessment instruments are then used to inform the frequency, intensity, and type of treatment services provided to the individual (Belenko et al., 2017). Of youth who enter the JJ system, only 68%-71% are screened, of which 48%-58% are identified as in need of SU treatment. However, only 15%-27% of the youth identified as in need of treatment were referred to SU treatment (Dennis et al., 2019).

Additionally, it is important that the SU screener results map on to clinical diagnostic tools (American Society of Addiction Medicine, 2014), such as the Diagnostic and Statistical Manual of Mental Disorders 5 (DSM-5; American Psychiatric Association, 2013). Juvenile justice agencies frequently use SU screening instruments that are not intended to diagnose SUDs (Vincent et al., 2012); nevertheless, the results of screening are used to inform SU treatment referrals. Therefore, it is important that JJ agencies use screening and assessment tools that are designed to diagnose SUDs, such as those that map on to DSM-5 diagnostic criteria. This is because the DSM-5 identifies several classes of substances (alcohol; cannabis; hallucinogens; inhalants; opioids; sedatives, hypnotics, or anxiolytics; stimulants; tobacco; other/unknown) for which a SUD may develop. According to the DSM-5, there are 11 symptoms that may diagnose a SUD, which can be grouped into the following: impaired control, social impairment, risky use, and pharmacological criteria (American Psychiatric Association, 2013). The presence of 0-1

symptom indicates no SUD, 2-3 symptoms indicates a mild SUD, 4-5 symptoms indicates a moderate SUD, and greater than six symptoms indicates a severe SUD (American Psychiatric Association, 2013). See Appendix A for the full list of SUD symptoms in the DSM-5.

Reliable and clinically useful screening instruments have been developed for identifying SUDs, but typically they have limited application in JJ agencies. The Adolescent Drug Involvement Scale (ADIS; Moberg & Hahn, 1991) is a 13-item self-report instrument designed to distinguish youth engaging in more problematic SU from youth experiencing minimal problems related to more minor SU. It takes approximately five minutes to complete and is free to access. However, the scale is not tied to any specific set of assessment criteria, so there are no guidelines for how to interpret the scale beyond the fact that higher total scores indicate more serious levels of drug involvement. The fact that this instrument was not designed to be a complex clinical instrument to diagnose SUDs and guide developments of treatment plans is problematic within JJ settings. The Drug Use Screening Inventory-Revised (DUSI-R; Tarter, 1990) is a 159-item self-report instrument that documents the level of involvement with several different drugs and quantifies severity of consequences associated with SU. It takes 20 to 40 minutes to complete, 20 minutes to score, and results are used to inform areas requiring comprehensive assessment. One useful facet of the DUSI-R is the Lie Scale which is used to determine validity of the youth's responses. This is particularly important given that JJ youth are known to sometimes be untruthful in reporting their SU (Harris et al., 2007). However, each paper questionnaire costs $5.00, and the software license for computerized administration and scoring costs $250.00 per year. The length of the DUSI-R and amount of time required to score it poses a problem in JJ settings because the screening process is designed to be quick and efficient. Although both the ADIS and DUSI-R are recommended by the Substance Abuse and

Mental Health Services Administration (SAMHSA) for use in the JJ system (Center for Substance Abuse Treatment, 2012), an alternative instrument is needed that overcomes these challenges.

One brief and free evidence-based screening instrument that maps on to the DSM-5 criteria for SUDs is the TCU DS 5 (Knight et al., 2014; Knight et al., 2018). This instrument is widely used by JJ agencies in the U.S. to identify youth with SUDs. The TCU DS 5 was originally developed based on DSM-IV-R SUD diagnostic criteria (Knight et al., 2018), and later updated to align with changes in the DSM-5, such as the none, mild, moderate, and severe SUD classifications. Previous work has validated the TCU DS 5 (Wiese et al., 2019) and shown similar SUD classification rates as the previous version of the TCU Drug Screen in a sample of justice-involved adolescents and adults (Knight et al., 2018). See Appendix B for the list of TCU DS 5 diagnostic items. Identifying youth with SUDs allows JJ agencies to then provide these youth with evidence-based services, which lowers their risk of future recidivism and involvement with the JJ system (Farabee et al., 2001).

There are three guiding principles, as outlined in the risk-need-responsivity (RNR) model, which are used to inform assessments and treatment decisions for justice-involved individuals who engage in SU (Andrews et al., 1990; Andrews et al., 2011). First, as reflected by the risk principle, it is important to determine an individual's likelihood of reoffending so that they are matched with the appropriate level of program intensity. An individual's risk level is static (i.e., cannot change as a result of intervention), and includes such things as race, gender, and number of prior offenses. Next, the need principle dictates that rehabilitation programs should directly target an individual's needs that are associated with their delinquent behaviors. Interventions are designed to target an individual's needs because they are dynamic (i.e., can be

altered). Needs include such things as substance use and criminal thinking. It is important to screen youth as they enter the JJ system to identify what their needs are so optimal treatment services can then be provided. Screening instruments must be thoroughly vetted for their validity and reliability to minimize the chances of individuals being referred to an inappropriate "dose" (too much or too little) or type of treatment (inappropriate intensity). . Finally, the responsivity principle emphasizes that interventions should be chosen based on how well a program fits with an individual's abilities and learning style. Currently, JJ agencies are utilizing the RNR framework to guide system reform efforts that seek to reduce recidivism rates while simultaneously improving public safety by directly targeting the unique needs of individuals (e.g., Schwartz et al., 1991; Seigle et al., 2014).

The primary objective of the current study is to compare identification and classification rates of the TCU DS 5 based on existing scoring procedures against IRT model results. Existing scoring procedures for the TCU DS 5 rely on CTT. Specifically, dichotomous answers (0 = *no*, 1 = *yes*) are summed together for the 11 diagnostic items, and that total is used to identify whether an individual does not have a SUD, or has a mild, moderate, or severe SUD. This is then compared to an IRT approach, wherein a computerized scoring algorithm is used to diagnose and classify SUDs. The current study explores the incremental value of implementing an IRT model.

Classical Test Theory

Classical test theory was developed by Spearman (1904) and is currently used to calculate scores for the TCU DS 5. The primary concern of CTT is reliability (Novick, 1966), which in psychology refers to the likelihood that an individual will get the same score on a measure when it is administered at different points in time. There are three important concepts underlying CTT: test score, error, and true score (Kean & Reilly, 2014). The test score, otherwise

known as the observed score, is a respondent's score determined by their responses to the items in a survey. For example, if a participant responded "yes" to four of the 11 items in the TCU DS 5, their test score would equal four. The amount of error in a measure cannot always be controlled for and refers to external stimuli that are affecting an individual's responses to the items in a measure. For example, if an individual is not paying attention or distracted while the TCU DS 5 is administered, they may mistakenly answer "yes" to an item that should have been answered "no." Finally, the true score is best thought of as a theoretical construct. It is an individual's score on a measure if all sources of error had been eliminated. While this is not possible, every person has a true score, which can be estimated using an equation that accounts for an individual's test score and error (i.e., Observed score = True score + Error; Magno, 2009). Therefore, CTT attempts to explain and reduce error, so that measures are more reliable and test scores more accurately align with true scores.

There are a few basic assumptions of CTT, which are often believed to be met. For example, an individual's true score is assumed to be uncorrelated with the measurement error (Lord & Novick, 2008). Additionally, just as the observed score is assumed to be the sum of the true score and the error (Magno, 2009), the variance of observed scores is also believed to equal the sum of the variances of the true scores and error (Lord & Novick, 2008). There are two additional assumptions required for CTT. The first is a linear relationship between the observed and true scores (Lord & Novick, 2008). This is necessary for CTT, as it allows for the linear combination (i.e., summing) of items to calculate scores. Further, the CTT model scales latent traits on an ordinal scale, which limits CTT compared to IRT, wherein scaling is typically at the interval level (Kean & Reilly, 2014). Interval scales are more powerful than ordinal scales and

typically allows for more detailed results (Kean & Reilly, 2014). The assumptions of CTT are not as stringent as those for IRT.

Item Response Theory

The primary objective of IRT is to measure and estimate where individuals fall on a latent continuum (i.e., a trait that cannot be directly measured, represented as theta or Θ) using stochastic models (Petrillo et al., 2015). Stochastic models are tools used to estimate distributions by randomly varying one or more data points and holding everything else constant. As mentioned previously, IRT requires more stringent assumptions compared to CTT. These assumptions include local independence, monotonicity, unidimensionality, and invariance. See below for a more detailed explanation of these assumptions, as well as how to test for them.

There are several aspects of IRT that factor into how the results are interpreted. In IRT, an item characteristic curve (ICC) is generated for each individual item, which depicts the probability of answering an item correctly based on an individual's ability (Kaplan & Saccuzzo, 1997). In other words, the ICC visually depicts the likelihood of endorsing an item based on where the individual falls on the latent trait being measured. The Rasch one-parameter logistic model is appropriate for modeling the probability of correct responses to dichotomous items, and assumes that the discriminations of all items are equal to one (Maier, 2001). An individual's position on the latent continuum changes is a function of their ability, and is determined by the sample's characteristics as well as the item parameters reflected in the ICCs (Anastasi & Urbina, 2002). Item parameters are unique for each individual item and determine the shape of the ICC. Item discrimination (a) is the parameter that determines the rate at which the probability of answering an item correctly changes based on individuals' ability levels (Magno, 2009). This parameter is reflected by the steepness of the curve at its steepest point. Steeper slopes reflect

items that are better at distinguishing between individuals. In contrast, items with low item discrimination values have more gradual curves and are not considered good items because they do not differentiate people well. Item difficulty (*b*) is the parameter that determines where an item falls along the ability scale (*x*-axis of the ICC; Magno, 2009). It is located at the ability point wherein 50% of respondents answer correctly. As the item becomes increasingly difficult, the curve shifts rightward, indicating a higher ability level for respondents that answer the item correctly (i.e., only people with high levels of the latent trait endorse that item). The final item parameter is the guessing parameter (*c*), which accounts for guessing on an item (Magno, 2009). Importantly, when a test consists of only dichotomous items, the one-parameter logistic model is used (Maier, 2001). This model only allows for individuals' abilities and item difficulties to vary. The discrimination parameter is fixed (i.e., they are all equal to one), and the guessing parameter is also not included. Therefore, since the first 11 items of the TCU DS 5 are dichotomous, the one-parameter logistic model is tested, so the item discrimination and guessing parameters are not analyzed. If the results showed that the discriminations of all of the items were not equal to one, then a two-parameter model would have been estimated.

There are four important assumptions for IRT. These include local independence, monotonicity, unidimensionality, and invariance. Local independence refers to the assumption that all the items in a test are not related to each other (McDonald, 1982). After the effect of the underlying trait is factored out, there should not be any relationship between two items (i.e., the residual covariance is zero). However, since the items in a test are assumed to only measure one latent trait (see unidimensionality assumption below), this assumption is never completely met. Research suggests that minor local dependence does not significantly affect IRT results (Wang & Wilson, 2005), so this assumption is assumed to be met.

Monotonicity refers to the assumption that as the trait level is increasing, the probability of a correct response also increases (Junker & Sijtsma, 2000). Using the TCU DS 5 as an example, as a person's SUD severity increases, the probability that they will respond "yes" to any of the first 11 items also increases. The relationship between trait and response is visually depicted in the category probability curves (CPC).

Unidimensionality refers to the assumption that the items included in the IRT analysis all measure the same latent variable, and that this latent trait is responsible for how the items in a test are responded to (Gordon et al., 2012). In the case of the TCU DS 5, all the items must measure SUD severity. This assumption can be tested via a factor analysis or principal component analysis (PCA). If the individual items load onto more than one factor (i.e., latent variable), then separate IRT analyses will be conducted for each of the factors.

The invariance assumption refers to the fact that an item's parameters can be estimated regardless of the sample characteristics within a population (Rupp & Zumbo, 2006). In the case of the TCU DS 5, the item parameters for the first item (*"Did you use larger amounts of drugs or use them for a longer time than you planned or intended?"*) estimated by an IRT model would not change regardless of respondent characteristics such as age, sex, or race. This assumption is tested via differential item functioning (DIF) analysis.

Comparisons between Classical Test Theory and Item Response Theory

There are some advantages of using CTT over IRT. First, CTT is far more common, and therefore familiar, among scientific audiences (Kean & Reilly, 2014). Accordingly, popular statistical packages often provide CTT statistical tests. The assumptions for CTT can be more easily met when compared to IRT, which makes CTT more widely applicable (Kean & Reilly,

2014). Additionally, IRT requires larger samples than CTT; specifically, IRT requires a sample size of between 100 to 150 participants (Kean et al., 2018).

While CTT has been used as the basis for the development of most tests in psychology and education (Kean & Reilly, 2014), there are advantages to using IRT instead of CTT. First, CTT primarily analyzes at the measure level (total score for the entire instrument), whereas IRT is more heavily focused on item-level analysis (Kean & Reilly, 2014). For this reason, the individual items that make up an instrument cannot be reduced or changed in any way after being validated using a CTT model, or else the new instrument must be completely re-assessed for validity and reliability. In contrast, since IRT analyzes each item individually for its validity, items can be removed while still maintaining validity. Additionally, IRT analyzes reliability of an instrument for each person individually, whereas CTT calculates reliability of an instrument overall for an entire population; consequently, using CTT to evaluate an instrument's reliability often results in longer surveys with more questions compared to using IRT (Kean & Reilly, 2014).

Current study

There are advantages and disadvantages to using both CTT and IRT. The current study compares TCU DS 5 results using both methods to explore whether the simple summative scoring scheme currently in place (CTT approach) is as good as the maximum likelihood estimate of the latent variable (true drug use severity) modeled by IRT. While IRT may account for more measurement error when predicting SUD severity, the added scoring burden of this approach may not make it worthwhile for agencies using the TCU DS 5.

While the current methodology for scoring the TCU DS 5 using a CTT approach has been established, it is possible that identification and classification of SUD severity scores may

benefit from use of a statistically-driven optimally weighted scoring algorithm generated via an IRT model. In other words, it is important to understand how the classifications of SUD correspond to the estimate of actual drug use severity. For example, in a test grading system using A's, B's, and C's, not everyone getting an A is actually equal in terms of their knowledge of the subject matter. Similarly, everyone classified as severe on the TCU DS 5 may not be the same in terms of the severity of their drug use problems. Of primary concern for the current study is whether an IRT approach can better assess SUD severity compared to the current CTT approach. Specifically, it is hypothesized that the use of an IRT model will be significantly better at determining TCU DS 5 total scores compared to the current CTT approach.

The use of an IRT model also allows for additional research questions to be addressed. Specifically, are some of the initial 11 diagnostic items better at differentiating between individuals with varying degrees of SUD severity? It is hypothesized that each of the initial 11 diagnostic items will differ in how well they discriminate between individuals with varying levels of SUD severity. Previous work has shown that DSM-5 criteria for SUD do not equally differentiate between SUD severity levels in a sample of adolescent heroin users (Yang et al., 2019). Since the TCU DS 5 is based on DSM-5 criteria for SUD, the same pattern of results is expected. Specifically, Item 5 ("*Did you get so high or sick from using drugs that it kept you from working, going to school, or caring for children?*") and Item 6 ("*Did you continue using drugs even when it led to social or interpersonal problems?*") will have the highest discrimination values (i.e., will be best at identifying severe SUDs). In contrast, Item 2 ("*Did you try to control or cut down your drug use but were unable to do it?*") and Item 11 (withdrawal criteria) will have the lowest discrimination values.

Exploratory correlational analyses will examine the relationship between TCU DS 5 total scores and responses to Items 12 through 17. It is hypothesized that most adolescents will identify alcohol and marijuana (Substance Abuse and Mental Health Services Administration, 2014) as their most used substance in the previous 12 months (Item 12), so there will most likely not be enough variation in responses to determine if certain classes of drugs significantly increase an individual's likelihood of developing a more severe SUD. Previous work has shown that, among JJ youth, marijuana is the most used substance, followed by alcohol, and then synthetic marijuana (Knight et al., 2018). For Item 13 ("*How often did you use each type of drug during the last 12 months?*"), it is hypothesized that responding more than "*Never*" to any substance in the previous 12 months will significantly increase an individual's likelihood of having a severe SUD. It is hypothesized that Item 14 ("*How many times before now have you ever been in a drug treatment program?*"), Item 15 ("*How serious do you think your drug problems are?*"), and Item 17 ("*How important is it for you to get drug treatment now?*") will not provide much information in differentiating SUD severity scores because (1) adolescents have likely not been in treatment before (Lipari et al., 2016) and (2) are not very good at identifying the severity of their own SU problem (Winters et al., 2014). In 2015, only 6.3% of adolescents aged 12 to 17 in need of SUD treatment received any type of treatment for their SU (Lipari et al., 2016). This lack of motivation to seek of treatment has been attributed to the limited perceived consequences adolescents have of their drug use and a lack of maturity that contributes to poor insights that a problem exists (Winters et al., 2014). However, Item 16 ("*During the last 12 months, how often did you inject drugs with a needle?*") will likely provide valuable information in terms of differentiating SUD severity among adolescents. Specifically,

any adolescent who reports injecting drugs with a needle more than "*Never*" will be at significantly increased risk of a severe SUD (Levy & Williams, 2016).

Section II: Method

Software

Item response theory analyses and some assumptions (monotonicity and invariance) were conducted using Winsteps Version 4.4.6 computer program (Linacre, 2019) and SAS Version 9.4 software (SAS Institute Inc., 2013). The unidimensionality assumption and correlational data were analyzed using SPSS Version 25.0 (IBM Corp., 2017).

Sample

Approval from TCU's Institution Review Board was obtained prior to study implementation. In the sample, a total of 312 male juveniles were recruited from two Midwestern correctional facilities. The TCU DS 5 was administered to all new admissions at intake between January and May 2016. The agencies and research center enacted a data sharing agreement prior to sharing de-identified data via a secure data service. Juvenile justice agency staff removed all personally identifiable information and assigned each youth a unique identifier prior to sending data to the research center. Participant ages ranged from 13 to 20 years old (M = 16.67, SD = 1.33). The sample was 62.8% black/African American, 23.1% white/Caucasian, and 14.1% Hispanic.

Instruments

***TCU Drug Screen 5*.** The TCU DS 5 is a valid, evidence-based screener for both adolescents and adults (Knight et al., 2014; Knight et al., 2018). It can be administered as either an independent self-report or in small groups (with a proctor reading each item aloud). There are 17 items in total, and respondents take approximately 5 min to complete the screen. For the first

11 items, respondents answer yes/no to a series of questions regarding their SU over the previous 12 months (before incarceration, if applicable). See Appendix B for a complete list of diagnostic items. Final scores are calculated by summing responses to the first 11 items, and scores range from 0 to 11 (this summative scoring procedure is based on DSM-5 SUD scoring recommendations; American Psychiatric Association, 2013). Note that Items 10 and 11 have two parts, and answering "yes" to either part corresponds to a score of 1. Specifically, Item 10 addresses the DSM-5 tolerance criteria (i.e., requiring more of a drug to feel the same effect as before, or the same amount of a drug resulting in less of an effect), and Item 11 addresses the withdrawal criteria (i.e., experiencing withdrawal symptoms if a drug is not taken, or continuing to take a drug to prevent experiencing withdrawal symptoms). A final score of 0-1 indicates no SUD, 2-3 indicates a mild SUD, 4-5 indicates a moderate SUD, and six or more indicates a severe SUD. Note that Items 12 through 17 are not included as part of the final TCU DS 5 total score; instead, these items are intended to provide additional information that may be helpful to inform treatment decisions. Item 12 lists 19 options respondents may select as the drug that caused the most serious problem during the last 12 months. Item 13 lists each of these drugs separately, and respondents indicate how often each type of drug was used during the last 12 months on a 5-point Likert scale (0 = *Never* to 4 = *Daily*). Item 14 asks respondents how many times they have been in a drug treatment program, with responses answered on a 5-point Likert scale (0 = *Never* to 4 = *4 or more times*). Item 15 asks respondents how serious they think their drug problems are, with responses on a 5-point Likert scale (0 = *Not at all* to 4 = *Extremely*). Item 16 asks how often they injected drugs with a needle over the previous 12 months, with responses on a 5-point Likert scale (0 = *Never* to 4 = *Daily*). Lastly, item 17 asks how important

it is for them to get drug treatment now, with responses on a 5-point Likert scale (0 = *Not at all* to 4 = *Extremely*).

Analytic plan

First, basic measurement statistics were calculated using a CTT approach, such as means and standard deviations for calculated total scores. Each individual's total score was calculated using the following steps: (1) One point is assigned to each "*yes*" response to Items 1 through 9. A "*yes*" response to either Item 10a or 10b, and Item 11a or 11b is assigned one point. (2) All one-point "*yes*" responses are summed for Items 1 through 11, such that total scores will range from 0 to 11. Next, total scores are converted into severity scores based on the following DSM-5 criteria. A score of 0-1 indicates no SUD. A score of 2-3 indicates a mild SUD. A score of 4-5 indicates a moderate SUD. A score of 6 or more indicates a severe SUD.

Next, assumptions for IRT were assessed to ensure none were violated. Local independence, or the assumption that the individual TCU DS 5 items were not related to each other was assumed. Category probability curves (CPC), which shows the relationship between the probability of a given category (i.e., item responses: yes vs. no) as a function of a person's ability (the construct being measured, in this case, severity of SUD) was calculated to test the assumption of monotonicity. Principal component analysis (PCA) was used to assess for unidimensionality. If the results of the PCA indicated that the first 11 items of the TCU DS 5 assess more than one dimension, or latent trait, then separate IRTs were performed on each dimension. Differential item functioning analysis was used to test for invariance. After ensuring that all assumptions had been met, the IRT analysis was conducted. Ultimately, the IRT analyses were designed to evaluate how well each item discriminated among individuals with varying levels of SUD. Most importantly, the simple summative scoring method generated by the CTT

analysis was compared to the IRT results wherein a statistically-driven optimally weighted scoring algorithm was generated to calculate TCU DS 5 scores, with the results depicted graphically via ICCs for each item.

A correlational analysis examined the relationship between individuals' ability levels computed via the IRT and total scores computed via CTT. A significant correlation would imply that the two methods of calculating TCU DS 5 scores were related to one another and therefore neither method was better than the other. A one-way analysis of variance (ANOVA) examined whether there was a significant difference between TCU DS 5 SUD severity diagnoses on the ability levels computed via the IRT. The R^2 statistic was calculated to inform the percentage of variation in ability scores accounted for by the TCU DS 5 SUD severity diagnoses. A high R^2 value indicates there is a high correspondence of being able to classify individuals on SUD severity between the two methods. The R^2 value coupled with the distributions of ability scores within SUD severity diagnoses (calculated using CTT) informs how much more additional information the IRT analysis is providing.

Further, some items may better differentiate among individuals with varying levels of SUD severity. Note that it is important for the 11 diagnostic items of the TCU DS 5 to have varying difficulty estimates, thus allowing the different items to not only differentiate severe SUDs, but also mild and moderate SUDs. As indicated above, based on previous research (Yang et al., 2019), it is hypothesized that Item 5 ("*Did you get so high or sick from using drugs that it kept you from working, going to school, or caring for children?*") and Item 6 ("*Did you continue using drugs even when it led to social or interpersonal problems?*") will have the highest discrimination values (i.e., will be best at identifying severe SUDs). In contrast, Item 2 ("*Did you*

try to control or cut down your drug use but were unable to do it?") and Item 11 (withdrawal criteria) will have the lowest discrimination values.

A correlational analysis examined the relationship between TCU DS 5 total scores and responses to Items 12 through 17. Based on findings from the 2013 National Survey on Drug Use and Health (Substance Abuse and Mental Health Services Administration, 2014), it is hypothesized that most adolescents will identify alcohol and marijuana as their most used substance in the previous 12 months (Item 12), so variation in responses might be too limited to determine if certain classes of drugs significantly increase an individual's likelihood of developing a more severe SUD. For Item 13, it is hypothesized that responding "*Daily*" use or using "*1-5 times per week*" of any substance in the previous 12 months will significantly increase an individual's likelihood of having a severe SUD. It is hypothesized that Item 14 ("*How many times before now have you ever been in a drug treatment program?*"), Item 15 ("*How serious do you think your drug problems are?*"), and Item 17 ("*How important is it for you to get drug treatment now?*") will not provide much information in differentiating SUD severity scores because adolescents have likely not been in treatment before (Lipari et al., 2016) and are not very good at identifying the severity of their own SU problem (Winters et al., 2014). However, it is hypothesized that Item 16 ("*During the last 12 months, how often did you inject drugs with a needle?*") will provide valuable information in terms of differentiating SUD severity among adolescents. Specifically, any adolescent who reports injecting drugs with a needle more than "*Never*" will be at significantly increased risk of a severe SUD.

Section III: Results

First, total scores for the TCU DS 5 were calculated using CTT. Total scores ranged from 0 to 11 (M = 3.24, SD = 3.89). These total scores were then converted into SUD severity scores (see Table 1). The majority of the sample (51.6%) did not have a diagnosable SUD.

Table 1

Frequency and Percentage of TCU Drug Screen 5 SUD Severity Diagnoses

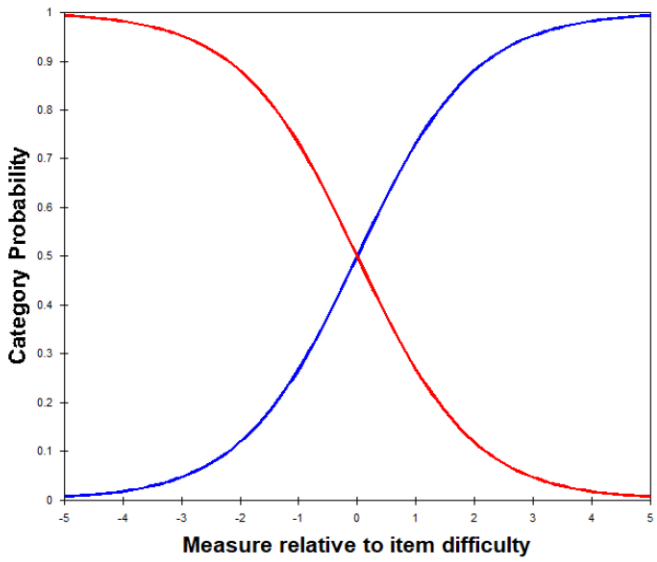
SUD Severity	Frequency	Percentage
None	161	51.6%
Mild	41	13.1%
Moderate	28	9.0%
Severe	82	26.3%

Note. SUD = substance use disorder.

To test for monotonicity, a CPC was generated for each of the 11 diagnostic items of the TCU DS 5. The *x*-axis represents the latent trait of SUD severity and the *y*-axis represents the probability of responding *"no*" or "*yes*" to the item. See Figure 1 for an example of a CPC for Item 1, which looks nearly identical to the CPC generated for the other items. Importantly, as an individual's SUD becomes increasingly severe, the probability of responding "*no*" to the items decreases, and the probability of responding "*yes*" increases. This confirms that the monotonicity assumption was met. Table 2 lists the average score individuals received who responded a certain way to each of the items. Importantly, responding "*yes*" to any of the 11 items resulted in a higher average score than responding "*no*."

Figure 1

Category Probability Curve (CPC) for Item 1 of the TCU Drug Screen 5



Note. The red line represents the probability of responding "*no*." The blue line represents the probability of responding *"yes*."

Table 2

Average Ability Scores for Each Response Category of the TCU Drug Screen 5

Item #	Response Option	Ability Mean
1	*No*	-2.74
	Yes	.99
2	*No*	-2.65
	Yes	1.22
3	*No*	-2.82
	Yes	.86
4	*No*	-2.81
	Yes	1.49

Table 2 (continued)

Item #	Response Option	Ability Mean
5	*No*	-2.52
	Yes	2.21
6	*No*	-2.74
	Yes	1.46
7	*No*	-2.63
	Yes	1.88
8	*No*	-2.49
	Yes	2.16
9	*No*	-2.65
	Yes	2.00
10	*No*	-3.04
	Yes	1.18
11	*No*	-2.72
	Yes	2.07

Note. Ability scores reflect average total scores (relative to each item) of persons who responded with each rating scale category.

A PCA was conducted to test for the unidimensionality assumption. Only one component was extracted, indicating that all items measure the same latent construct. The first and only factor had an eigenvalue of 6.80, and accounted for most (61.60%) of the variance in the data. With a cutoff of 0.40 for inclusion of a variable in the factor (Matsunaga, 2010), all 11 items loaded onto this single factor. See Table 3 for each item's factor loading.

Table 3

Item Fit Statistics

Item #	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	PTMZ *R* Corr.	PCA Loading
1	1.40	3.77	1.75	3.90	0.71	0.67
2	1.36	3.26	1.57	3.14	0.71	0.69
3	1.39	3.82	1.83	3.89	0.71	0.66
4	0.88	-1.24	0.92	-0.45	0.80	0.82
5	0.83	-1.50	0.58	-1.97	0.79	0.82
6	1.04	0.47	1.06	0.45	0.77	0.78
7	0.87	-1.14	0.84	-0.83	0.79	0.82
8	0.89	-0.89	0.89	-0.37	0.77	0.80
9	0.74	-2.50	0.59	-2.42	0.81	0.85
10	0.83	-1.97	0.67	-2.02	0.82	0.82
11	0.58	-4.52	0.42	-4.00	0.84	0.89

Note. MNSQ = mean-square; ZSTD = standardized weighted (infit) and unweighted (outfit) mean-squared fit statistics; PTMZ = point-measure correlation; PCA = principal component analysis.

Table 3 also contains fit statistics. These fit statistics include infit/outfit mean-squared statistics, standardized weighted (infit) and unweighted (outfit) mean-squared statistics, and point-measure correlations. Infit and outfit mean-squared statistics indicate how well the items fit with the overall model, and both are influenced by outlier cases. While infit is most affected by outlier responses to items that best target the average person, outfit is most affected by outlier responses to items that best target persons that fall on the upper and lower extremes of the latent continuum (Kean et al., 2018). In both instances, misfit of items is indicated by values less than 0.6 or more than 1.4 (Kean et al., 2018). Accordingly, the infit and outfit mean-squared statistics

identify Items 1, 2, 3, 9, and 11 as being potentially misfit. These items correspond with the larger/longer, quit/control, time spent, physical/psychological problems, and withdrawal criteria of the DSM-5, respectively. Note that Item 9 is misfit, with an outfit mean-square statistic of 0.59. Similarly, the standardized weighted and unweighted mean-squared statistics indicate misfit by values less than -2.0 or greater than 2.0 (Kean et al., 2018). The thresholds indicate again that Items 1, 2, 3, 9, and 11 are misfit, implying that they may be measuring a different construct than the other items, are poor quality items, or there are errors in the data quality (Kean et al., 2018). The pattern of results for infit/outfit statistics indicate that responses to Items 1, 2, and 3 are highly unpredictable, whereas Items 9 and 11 are highly predictable. While the PCA results verify that all of the items are measuring the same underlying construct, the TCU DS 5 may benefit from adjusting the way these items are worded. Predictable items (9 and 11) could be replaced with more efficient items, although even items with very low mean-squared values still add a little bit of new and useful information (Martin-Löf, 1974), and therefore should be retained as they are to maintain consistency with DSM-5 criteria. Additionally, the infit and outfit values for Items 9 and 11 may be affected by the unpredictable response patterns of Items 1, 2, and 3. It is recommended that items with very unpredictable response patterns be removed or re-worded (Martin-Löf, 1974). It is important to note that since the mean-squared values for Items 1, 2 and 3 are less than 2.0, they are not too misfit that they are distorting or degrading the measurement system (Gustafsson, 1980). Therefore, it is recommended that these items be re-worded but not removed entirely. The point-measure correlation is a Pearson correlation between the individual items and how the sample is responding to the other items in the model. Point-measure correlation values less than 0.4 indicate misfit (Kean et al., 2018). These values are all greater than 0.4, indicating individuals are responding similarly to all the items.

Given that item difficulty indicates the average ability levels of persons wherein 50% of the sample is correctly endorsing the item, higher difficulty values indicate that the respondents answering "*yes*" to the item are likely to have more severe SUDs. Lower values indicate that individuals with more mild SUDs are also answering "*yes*" to the item. See Table 4 for the difficulty estimates of each item. Figure 2 depicts ICCs for Items 8 and 10. Item 8 has the highest difficulty estimate, whereas Item 10 has the lowest. Thus, individuals who use drugs that put themselves or others in physical danger (Item 8) likely have more severe SUDs. In contrast, individuals with more mild forms of SUD are still likely to experience tolerance symptoms (Item 10). Stated differently, tolerance symptoms seem to develop early on after someone first starts engaging in SU, and therefore can act as an early warning signal that an adolescent is on the path to developing a SUD.

Table 4

Item Difficulty Estimates and Standard Errors

Item #	Difficulty Estimate	Standard Error
1	0.52	0.10
2	0.62	0.10
3	0.44	0.10
4	0.53	0.08
5	0.80	0.09
6	0.58	0.09
7	0.70	0.09
8	0.82	0.09
9	0.68	0.08

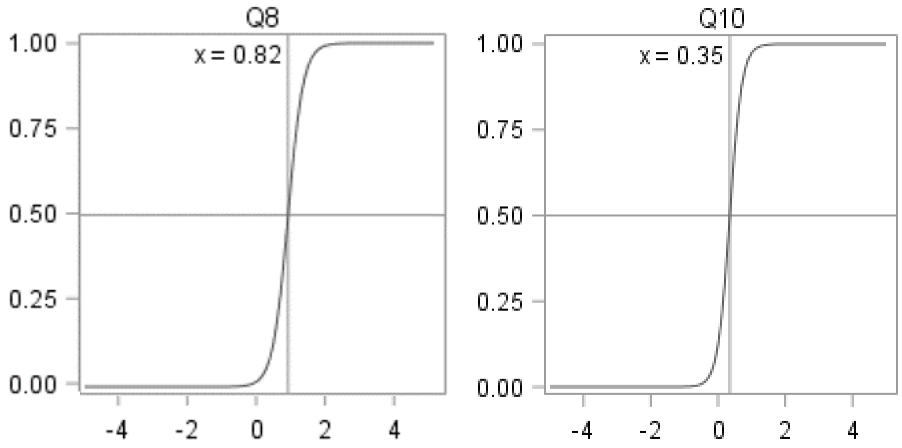
Table 4 (continued)

Item #	Difficulty Estimate	Standard Error
10	0.35	0.08
11	0.63	0.07

Note. Difficulty estimates reflect the average ability levels wherein 50% of respondents are correctly endorsing the item.

Figure 2

Item Characteristic Curves (ICC) for Items 8 and 10 of the TCU Drug Screen 5



Note. The *y*-axis is the probability of responding "*yes*" to the item. The *x*-axis represents ability level.

A one-way between-subjects analysis of variance (ANOVA) was performed looking at CTT severity diagnoses (none, mild, moderate, vs. severe) on IRT ability scores. There was a significant effect of SUD severity diagnosis, $F(3, 308) = 819.77$, $p \leqslant .001$, $R^2 = 0.88$, $R^2_{Adjusted} = 0.88$, 95% CI [0.51, 0.55]. All groups were significantly different from each other, ps $\leqslant .001$. See Table 5 for descriptive statistics. Nearly all (88%) of the variance in the ability scores are

explained by CTT severity diagnoses. In other words, there is a high correspondence between the classification system underlying the TCU DS 5 and the estimate of true drug use severity.

Table 5

Ability Scores by SUD Severity Diagnosis

Severity Diagnosis	Descriptive Statistics
None	-0.49 (0.02)
Mild	0.24 (0.04)
Moderate	0.51 (0.05)
Severe	1.10 (0.03)

Note. Values are means and standard errors (in parentheses).

Exploratory analyses examined how IRT ability scores and CTT total scores were related to the supplementary items (Items 13-17) of the TCU DS 5. Just as how the majority of the sample did not have a diagnosed SUD, most people (47%) reported that no drug caused them the most problems (Item 12). Marijuana was the drug most often (35%) identified as causing respondents the most serious problem. See Table 6 for the full list of drugs in the TCU DS 5. Given the uniformity in responses to Item 12, no additional analyses were conducted to examine this item. Correlational analyses examined the relationship between supplementary Items 13-17 and TCU DS 5 total scores (via CTT) and ability scores (via IRT). All correlations were significant for both total scores and ability scores except for the following items: 13c, 13d, 13f, 13g, 13k, 13o, and 14. Recall that Item 13 lists several types of drugs and asks how often each was used during the last 12 months. Ability scores were significantly related to Item 13c ("*cannabinoids—hashish (hash)*") and Item 13d ("*synthetic marijuana (K2/spice)*"), *p*s $\leqslant$ .042, but total scores were not related to either, *p*s $\geqslant$ .063. The opposite pattern of results was found

for Item 13k ("*bath salts (synthetic cathinones*"), which was not significantly related to ability scores, $p = .141$, but was significantly related to total scores, $p = .048$. Neither ability scores nor total scores were related to Items 13f ("*opioids—opium (tar)*"), 13g ("*stimulants—powder cocaine (coke)*"), 13o ("*inhalants—solvents (paint thinner)*"), or 14 ("*How many times before now have you ever been in a drug treatment program?*"), $ps \geqslant .144$. See Table 7 for correlations between supplementary items and total scores.

Table 6

Descriptive Statistics for Supplementary Items 12 and 13

Type of Drug	# of People Who Say This Drug Causes Them the Most Serious Problem	Average of How Often Drug is Used (*M*)	*SD*
None	147	--	--
Alcohol	15	0.87	1.19
Cannabinoids – Marijuana (weed)	108	2.47	2.42
Cannabinoids – Hashish (hash)	1	0.14	0.57
Synthetic Marijuana (K2/spice)	10	0.29	0.91
Opioids – Heroin (smack)	0	0.03	0.23
Opioids – Opium (tar)	1	0.03	0.26
Stimulants – Powder Cocaine (coke)	5	0.16	0.61
Stimulants – Crack Cocaine (rock)	0	0.04	0.24
Stimulants – Amphetamines (speed)	3	0.17	0.69
Stimulants – Methamphetamine (meth)	8	0.13	0.64
Bath Salts (synthetic cathinones)	0	0.02	0.20
Club Drugs – MDMA/GHB/Rohypnol (ecstasy)	4	0.24	0.77
Dissociative Drugs – Ketamine/PCP (Special K)	2	0.06	0.41
Hallucinogens – LSD/Mushrooms (acid)	0	0.12	0.50

Table 6 (continued)

Type of Drug	# of People Who Say This Drug Causes Them the Most Serious Problem	Average of How Often Drug is Used (*M*)	*SD*
Inhalants – Solvents (paint thinner)	0	0.06	0.39
Prescription Medications – Depressants	5	0.36	0.96
Prescription Medications – Stimulants	1	0.07	0.43
Prescription Medications – Opioid Pain Relievers	0	0.25	0.82
Other	2	0.25	0.78

Note. M = mean; *SD* = standard deviation.

Table 7

Relationships Between Supplementary Items, Ability Estimates and Total Scores

Item #	Ability Estimates (IRT)	Total Scores (CTT)
13a	0.29**	0.19**
13b	0.18*	0.11*
13c	0.16*	0.11
13d	0.12*	0.10
13e	0.12*	.012*
13f	0.06	0.06
13g	0.08	0.08
13h	0.14*	0.15*
13i	0.19**	0.20**
13j	0.15*	0.19**
13k	0.08	0.11*
13l	0.19**	0.17*
13m	0.12*	0.12*
13n	0.12*	0.11*
13o	0.08	0.08

Table 7 (continued)

Item #	Ability Estimates (IRT)	Total Scores (CTT)
13p	0.21**	0.21**
13q	0.12*	0.12*
13r	0.14*	0.14*
13s	0.13*	0.11*
14	0.05	0.03
15	0.34**	0.31**
16	0.11*	0.11*
17	0.33**	0.32**

Note. Numbers represent Pearson correlations. IRT = Item Response Theory; CTT = Classical Test Theory. **$p \leqslant .001$, *$p < .05$.

Section IV: Discussion

Overall, the results suggest that while IRT provides valuable additional information on SUD severity, it is likely too burdensome to be the standard scoring method of the TCU DS 5. The screener is designed to be administered and manually scored by justice staff, such as probation and parole officers, who may not be familiar with, or capable of, the computation of IRT analysis. Being able to quickly score by summing up "*yes*" responses provides an easy and quick way to identify youth who are at-risk of having a SUD at intake. The screener results can then be used to refer at-risk youth to receive a full assessment, which can dive deeper into an adolescent's SU to develop a more holistic understanding of their individual needs.

The IRT identified five items (Items 1, 2, 3, 9, and 11) as being misfit. While responses to Items 9 and 11 were highly predictable, responses to Items 1, 2, and 3 were much less predictable. Even items with very low mean-squared values still add a little bit of new and useful information (Martin-Löf, 1974), and therefore Items 9 and 11 should be retained as they are to maintain consistency with DSM-5 criteria. However, Items 1, 2, and 3 should be re-worded. Item

1 asks whether individuals used larger amounts of drugs or used them for a longer time than they planned or intended. This may be problematic because an individual could respond "*no*" because they had no intention of cutting back their drug use, even if they have a severe SUD. Previous work has suggested that requiring an intention to cut back SU is problematic (Substance Abuse and Mental Health Services Administration, 2016). This item could be changed to instead to ask whether an individual has increased the amount or frequency of their drug use since when they first started using, without specifying any intention to cut back their use.

Item 2 asks whether individuals tried to control or cut down their drug use but were unable to. Again, as with Item 1, an individual with a severe SUD may have never tried to limit their drug use, so therefore would respond "*no*" to this question. This item could be re-worded such that the word "try" is removed; for example, "Have you been unsuccessful at controlling or cutting down your drug use?"

Item 3 asks if an individual spent a lot of time getting drugs, using them, or recovering from their use. This may be problematic because an individual may respond "*yes*" to spending a lot of time getting drugs simply because they are more difficult to get in their area, and not because they have a more severe SUD. In contrast, an individual who spends a lot of time using drugs would also respond "*yes*" to this question, but their answer would reflect a more severe SUD. This item should be considered for re-wording so that it asks only about whether an individual spent a lot of time using drugs or recovering from using them, and not about how much time they spent trying to get them. For example, "Did you spend a lot of time using drugs or recovering from their use?"

The IRT also provided insight into how people with different severities of SUDs respond to the items. Item 10 (tolerance criteria) had the lowest difficulty estimate, which means that

tolerance is likely the first SUD symptom to appear in individuals with mild SUDs. If an individual endorses this item but not the others, they are likely a good target for a preventive intervention to try to curb their SU before it gets worse. In contrast, Items 5, 7, 8, and 9 had the highest difficulty estimates. Endorsing these items means that individuals have severe SUDs. Adolescents who report getting so high or sick from their drug use that they do not go to work or school (Item 5), spending less time at work, school, or with friends because of their drug use (Item 7), using drugs that put themselves or others in physical danger (Item 8), or continuing to use drugs despite it causing physical or psychological problems (Item 9) should likely be referred for high intensity treatment interventions for their SU.

The majority of the sample reported that no drug caused them the most serious problem, followed by marijuana. While the ability scores computed via the IRT analysis were related to how often hashish and synthetic marijuana was used, total scores computed via CTT were not. The opposite was true for bath salts, for which only CTT total scores were related. Neither method of computing total scores was related to how often opium, powder cocaine, or inhalants were used. Bath salts, opium, cocaine, and inhalants are all relatively "harder" drugs compared to others on the list and, with the exception of powder cocaine, are not commonly used among the sample. Reporting using these drugs at all, even only a few times, could indicate a potential severe SUD. This may explain the lack of relationship these variables have with total scores.

Finally, total scores were assessed for their relationship with remaining supplementary Items 14-17. How many times individuals have been in a drug treatment program (Item 14) was, as expected, not related to total scores. Regardless of how severe an adolescent's SUD is, they likely have not been to many drug treatment programs simply due to their young age. This is consistent with previous findings that only 6.3% of adolescents aged 12 to 17 in need of SUD

treatment received any type of treatment for their SU (Lipari et al., 2016). Unexpectedly, Items 15 ("*How serious do you think your drug problems are?*") and 17 ("*How important is it for you to get drug treatment now?*") were related to total scores. It was hypothesized that these items would not be related to total scores because adolescents would not have much insight into the severity of their SU (Winters et al., 2014); however, this was not the case. Adolescents in this study appear to have some level of awareness of how serious their SU was and how important it was to receive treatment. Lastly, Item 16 ("*During the last 12 months, how often did you inject drugs with a needle?*"), as hypothesized, was related to total scores. In total, only five respondents said they injected drugs more than "*never*." Of these five individuals, four had a severe SUD and one had no SUD. This is consistent with previous findings that certain drug use patterns, such as injection drug use, indicate acute risk of harm that warrants immediate attention (Levy & Williams, 2016). Reporting injecting drugs at all should require a referral to intensive treatment, regardless of SUD severity scores.

There are limitations to this study that should be addressed. First, the sample was all male juveniles, so the results do not necessarily extend to females and adults. Additionally, these adolescents were incarcerated at the time the TCU DS 5 was administered, so the results may not generalize to individuals in a non-justice sample. This may have also affected how respondents answered the questions. Because they were incarcerated, they may have felt answering truthfully could negatively affect their supervision requirements, despite agency staff informing them that truthful responses would not negatively impact release requirements. The misfit of Items 1, 2, and 3 poses a potential problem for IRT. The inflated infit/outfit statistics for these items implies response patterns are highly unpredictable. The lack of fit between these items and the model could have skewed some estimates generated by the IRT analysis. The retrospective nature of the

data set may have been affected by recall bias. The fact that these adolescents had SUDs may have introduced a systematic error wherein they do not accurately remember previous events or unintentionally omitted important details. If recall bias was present, it may have led to different results than if this study was conducted prospectively. Finally, the supplementary items were only analyzed using correlational analyses, so these results should be interpreted with caution.

Future work should examine whether the suggested changes to Items 1, 2, and 3 results in improved fitness with the overall model. Additionally, another question should be added to check for the veracity of statements and make sure respondents are carefully reading each question. This would be helpful to easily identify which respondents to omit from the analysis. Additionally, future work should attempt to replicate these findings in a more diverse sample that includes females and adults. One helpful feature of IRT is the ability to conduct item bias analyses. This would ensure that diverse groups of individuals (i.e., females, adults, non-incarcerated individuals) are responding similarly to each of the items.

In conclusion, for field applications, the traditional way of scoring the TCU DS 5 appears to be worth continuing given the accuracy of scoring results and ease of administration, but the IRT analysis did provide insights into the TCU DS 5 that would not have been apparent using typical CTT analyses. For example, it identified additional diagnostic items that were endorsed by individuals who have a severe SUD. These items may serve as strong indicators that the individual potentially would benefit from intensive SUD treatment. Agencies who administer the TCU DS 5 should be aware that adolescents who endorse these items are most at risk.

Appendix A: Diagnostic and Statistical Manual of Mental Disorders 5 Substance Use Disorder Criteria

A. A problematic pattern of substance use leading to clinically significant impairment or distress, as manifested by at least two of the following, occurring within a 12-month period:

1. The substance is often taken in larger amounts or over a longer period than was intended.

2. There is a persistent desire or unsuccessful efforts to cut down or control substance use.

3. A great deal of time is spent in activities necessary to obtain the substance, use the substance, or recover from its effects.

4. Craving, or a strong desire or urge to use the substance.

5. Recurrent substance use resulting in a failure to fulfill major role obligations at work, school, or home.

6. Continued substance use despite having persistent or recurrent social or interpersonal problems caused or exacerbated by the effects of the substance.

7. Important social, occupational, or recreational activities are given up or reduced because of substance use.

8. Recurrent substance use in situations in which it is physically hazardous.

9. Substance use is continued despite knowledge of having a persistent or recurrent physical or psychological problem that is likely to have been caused or exacerbated by the substance.

10. Tolerance, as defined by either of the following:

    a. A need for markedly increased amounts of the substance to achieve intoxication or desired effect.

    b. A markedly diminished effect with continued use of the same amount of the substance.

11. Withdrawal, as manifested by either of the following:

    a. The characteristic withdrawal syndrome for the substance.

    b. The substance (or a closely related substance) is taken to relieve or avoid withdrawal symptoms.

Appendix B: TCU Drug Screen 5 Diagnostic Items for the Item Response Theory Model

During the last 12 months (before being locked up, if applicable) –

1. Did you use larger amounts of drugs or use them for a longer time than you planned or intended?
2. Did you try to control or cut down of your drug use but were unable to do it?
3. Did you spend a lot of time getting drugs, using them, or recovering from their use?
4. Did you have a strong desire or urge to use drugs?
5. Did you get so high or sick from using drugs that it kept you from working, going to school, or caring for children?
6. Did you continue using drugs even when it led to social or interpersonal problems?
7. Did you spend less time at work, school, or with friends because of your drug use?
8. Did you use drugs that put you or others in physical danger?
9. Did you continue using drugs even when it was causing you physical or psychological problems?

10a. Did you need to increase the amount of a drug you were taking so that you could get the same effects as before?

10b. Did using the same amount of a drug lead to it having less of an effect as it did before?

11a. Did you get sick or have withdrawal symptoms when you quit or missed taking a drug?

11b. Did you ever keep taking a drug to relieve or avoid getting sick or having withdrawal symptoms?

REFERENCES

American Society of Addiction Medicine. (2014). *The ASAM standards of care for the addiction specialist physician.* American Society of Addiction Medicine, Inc.

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*. Arlington, VA: American Psychiatric Association.

Anastasi, A., & Urbina, S. (2002). *Psychological testing.* Prentice Hall: New York.

Andrews, D. A., Bonta, J., & Hoge, R. D. (1990). Classification for effective rehabilitation: Rediscovering psychology. *Criminal Justice and Behavior, 17,* 19-52.

Andrews, D. A., Bonta, J., & Wormith, J. S. (2011). The risk-need-responsivity (RNR) model: Does adding the good lives model contribute to effective crime prevention? *Criminal Justice and Behavior, 38*(7), 735-755.

Belenko, S., Knight, D., Wasserman, G. A., Dennis, M. L., Wiley, T., Taxman, F. S.,… Sales, J. (2017). The Juvenile Justice Behavioral Health Services Cascade: A new framework for measuring unmet substance use treatment services needs among adolescent offenders. *Journal of Substance Abuse Treatment, 74,* 80-91. doi: 10.1016/j.jsat.2016.12.012

Center for Behavioral Health and Quality. (2016). *2015 National Survey on Drug Use and Health: Detailed Tables.* Rockville, MD: Substance Abuse and Mental Health Services Administration.

Center for Substance Abuse Treatment. (2012). *Screening and Assessing Adolescents for Substance Use Disorders.* Treatment Improvement Protocol (TIP) Series, No. 31. HHS Publication No. (SMA) 12-4079. Rockville, MD: Substance Abuse and Mental Health Services Administration.

Dennis, M. L., White, M. K., & Ives, M. L. (2009). In C. G. Leukefeld, T. P. Gullotta, & M. Staton-Tindall (Eds.), *Issues in children's and families' lives. Adolescent substance abuse: Evidence-based approaches to prevention and treatment. Individual characteristics and needs associated with substance misuse of adolescents and young adults in addiction treatment* (pp. 45-72). New York, NY: Springer Science + Business Media.

Dennis, M. L., Smith, C. N., Belenko, S., Knight, D., McReynolds, L., Rowan, G., Dembo, R., DiClemente, R., Robertson, A., & Wiley, T. (2019). Operationalizing a behavioral health services cascade of care model: Lessons learned from a 33-site implementation in juvenile justice community supervision. *Federal Probation, 83*(2), 52-64.

Donenberg, G. R., Emerson, E., Mackesy-Amiti, M. E., & Udell, W. (2015). HIV-risk with juvenile offender on probation. *Journal of Child and Family Studies, 24*(6), 1672-1674.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

Farabee, D., Shen, H., Hser, Y., Grella, C. E., & Anglin, M. D. (2001). The effect of drug treatment on criminal behavior among adolescents in DATOS-A. *Journal of Adolescent Research, 16,* 679-696.

Gordon, D., Howe, L. D., Galobardes, B., Matijasevich, A., Johnston, D., Onwujekwe, O., … & Hargreaves, J. R. (2012). Authors' response to: Alternatives to principal components analysis to derive asset-based indices to measure socio-economic position in low- and middle-income countries: The case for multiple correspondence analysis. *International Journal of Epidemiology, 41*(4), 1209-1210.

Gustafsson, J. E. (1980). Testing and obtaining fit of data to the Rasch model. *British Journal of Mathematical and Statistical Psychology, 33*(2), 205-233.

Harris, K. M., Griffin, B. A., McCaffrey, D. F., & Morra, A. R. (2007). Inconsistencies in self-reported drug use by adolescents in substance abuse treatment: Implications for outcomes and performance measurement. *Journal of Substance Abuse Treatment, 34*(3), 347-355. doi: 10.1016/j.jsat.2007.05.004

Henggeler, S. W., Clingempeel, W. G., Brondino, M. J., & Pickrel, S. G. (2002). Four-year followup of multisystemic therapy with substance-abusing and substance dependent juvenile offender. *Journal of the American Academy of Child and Adolescent Psychiatry, 41*(7), 868-874.

Hubley, A. M. (2002). Adaptive and tailored testing (including IRT and non-IRT application). *Encyclopedia of Psychological assessment*, 9-13.

IBM Corp. Released 2017. IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp.

Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement, 24*(1), 65-81.

Kaplan, R. M., & Saccuzo, D. P. (1997). *Psychological testing: Principles, applications and issues*. Pacific Grove: Brooks Cole Pub. Company.

Kean, J. K., Brodke, D. S., Biber, J., & Gross, P. (2018). An introduction to item response theory and Rasch analysis of the Eating Assessment Toll (EAT-10). *Brain Impairment, 19*(1), 91-102.

Kean, J., & Reilly, J. (2014). Classical test theory. *Handbook for Clinical Research: Design, Statistics and Implementation*, 192-194.

Knight, D. K., Becan, J. E., Landrum, B., Joe, G. W., & Flynn, P. M. (2014). Screening and assessment tools for measuring adolescent client needs and functioning in substance abuse treatment. *Substance Use & Misuse, 49*(7), 902-918. doi: 10.3109/10826084.2014.891617

Knight, D. K., Blue, T. R., Flynn, P. M., & Knight, K. (2018). The TCU drug screen 5: Identifying justice-involved individuals with substance use disorders. *Journal of Offender Rehabilitation, 57*(8), 525-537.

Levy, S. J., & Williams, J. F. (2016). Substance use screening, brief intervention, and referral to treatment. *Pediatrics, 138*(1), e20161211.

Lipari, R. N., Park-Lee, E., & Van Horn, S. (2016). *The CBHSQ Report: American's need for and receipt of substance use treatment in 2015.* Rockville, MD: Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration.

Linacre, J. M. (2019). Winsteps® Rasch measurement computer program. Beaverton, Oregon: Winsteps.com

Lord, F. M., & Novick, M. R. (2008). *Statistical Theories of Mental Test Scores*. Information Age Publishing Inc.

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *The International Journal of Educational and Psychological Assessment, 1*(1), 1-11.

Maier, K. S. (2001). A Rasch hierarchical measurement model. *Journal of Educational and Behavioral Statistics, 26*, 307-331.

Martin-Löf, P. (1974). The notion of redundancy and its use as a quantitative measure of the discrepancy between a statistical hypothesis and a set of observational data [with Discussion]. *Scandinavian Journal of Statistics*, 3-18.

Matsunaga, M. (2010). How to factor-analyze your data right: Do's, don'ts, and how-to's. *International Journal of Psychological Research, 3*(1), 97-110.

McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement, 6*, 379-396.

Meade, A. W., & Lautenschlager, G. J. (2004). Same question, different answers: CFA and two IRT approaches to measurement invariance. In *19th Annual Conference of the Society for Industrial and Organizational Psychology* (Vol. 1).

Moberg, D. P., & Hahn, L. (1991). The Adolescent Drug Involvement Scale. *Journal of Adolescent Chemical Dependency, 2*(1), 75-88.

Muthén, B., & Lehman, J. (1985). Multiple group IRT modeling: Applications to item bias analysis. *Journal of Educational Statistics, 10*(2), 133-142.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3*(1), 1-18.

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement, 66*(1), 63-84.

SAS Institute Inc. (2013). *SAS® 9.4*. Cary, NC: SAS Institute.

Schwartz, I. M., Barton, W. H., & Orlando, F. (1991). Keeping kids out of secure detention: The misuse of juvenile detention has a profound impact on child welfare. *Public Welfare, 49*(2), 20-26.

Seigle, E., Walsh, N., & Weber, J. (2014). *Core principles for reducing recidivism and improving other outcomes for youth in the juvenile justice system*. Council of State Governments.

Substance Abuse and Mental Health Services Administration. (2014). *Results from the 2013 National Survey on Drug Use and Health: Summary of National Findings*, NSDUH Series H-48, HHS Publication No. (SMA) 14-4863. Rockville, MD: Substance Abuse and Mental Health Services Administration.

Substance Abuse and Mental Health Services Administration. (2016). *Impact of the DSM-IV to DSM-5 Changes on the National Survey on Drug Use and Health* [Internet]. Rockville, MD: Substance Abuse and Mental Health Services Administration.

Spearman, C. (1904). "General intelligence," objectively determined and measured. *The American Journal of Psychology, 15*(2), 201-292. doi: 10.2307/1412107

Tapia, M., McCoy, H., & Tucker, L. (2016). Suicidal ideation in juvenile arrestees: Exploring legal and temporal factors. *Youth Violence and Juvenile Justice, 14*(4), 468-483.

Tarter, R. (1990). Evaluation and treatment of adolescent substance use: A decision tree method. *American Journal of Drug and Alcohol Abuse, 16*, 1-46.

Vincent, G. M., Guy, L. S., & Grisso, T. (2012). Risk assessment in juvenile justice: A guidebook for implementation. *Systems and Psychosocial Advances Research Center Publications and Presentations,* 573. https://escholarship.umassmed.edu/psych_cmhsr/573

Wang, W., & Wilson, M. (2005). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement, 29*(4), 296-318.

Wiese, A. L., Blue, T. R., Knight, D. K., & Knight, K. (2019). The validity of TCU Drug Screen 5 for identifying substance use disorders among justice-involved youth. *Federal Probation, 83*, 65-70.

Winters, K. C., Tanner-Smith, E. E., Bresani, E., & Meyers, K. (2014). Current advances in the treatment of adolescent drug use. *Adolescent Health, Medicine and Therapeutics, 5*, 199-210. doi: 10.2147/AHMT.S48053

Yang, H., Chen, F., Liu, X., & Xin, T. (2019). An item response theory analysis of DSM-5 heroin use disorder in a clinical sample of Chinese adolescents. *Frontiers in Psychology, 10*, 2209. doi: 10.3389/fpsyg.2019.02209

VITA

Personal Background	Amanda Lee Wiese Fort Worth, Texas Daughter of Christopher Kurth and Barbara Lee Wiese
Education	Bachelor of Science, Psychology, California Lutheran University, Thousand Oaks, California, 2015 Master of Science, Psychological Sciences, University of Texas at Dallas Richardson, Texas, 2018
Experience	Graduate student research assistant, Center for BrainHealth, 2016-2018 Research assistant, University of Texas Southwestern, 2017-2018 Graduate student research assistant, Texas Christian University Fort Worth, 2018-present Teaching Assistantship, Texas Christian University, 2020-present
Professional Memberships	Psi Chi International Honor Society Golden Key International Honour Society American Psychological Association

ABSTRACT

ANALYSES OF THE TCU DRUG SCREEN 5:

USING AN ITEM RESPONSE THEORY MODEL

WITH A SAMPLE OF JUVENILE JUSTICE YOUTH

by Amanda Lee Wiese, M.S., 2020
College of Science and Engineering
Texas Christian University

Thesis Advisor: Kevin Knight, Ph.D., Director of the Institute of Behavioral Research

Cathy Cox, Ph.D., Associate Professor and Director of Graduate Studies

Danica Kalling Knight, Ph.D., Associate Professor

George Joe, ED.D., Associate Director for Process and Outcome Studies

Mary Hargis, Ph.D., Assistant Professor

It is important to identify youth who have a substance use disorder (SUD) when they enter the juvenile justice (JJ) system using a screener such as the TCU Drug Screen 5 (TCU DS 5), so that necessary treatments can be provided to them. While the TCU DS 5 is a valid, evidence-based screener, the use of an item response theory (IRT) model may better differentiate among mild, moderate, and severe forms of SUD. The current study analyzes the feasibility and incremental value gained in using an IRT model to compute total TCU DS 5 scores compared to its current scoring methodology. The results reveal that while IRT may not be worthwhile as the standard method of scoring, there are benefits to using IRT to assess the validity and value of individual items in a screening instrument.