



A Virtual Standardized Patient–Based Assessment Tool to Evaluate Psychiatric Residents’ Psychopharmacology Proficiency

Jeffrey J. Rakofsky¹ · Thomas B. Talbot² · Boadie W. Dunlop¹

Received: 6 February 2020 / Accepted: 5 July 2020 / Published online: 17 July 2020
© Academic Psychiatry 2020

Abstract

Objectives A virtual standardized patient-based assessment simulator was developed to address biases and practical limitations in existing methods for evaluating residents’ proficiency in psychopharmacological knowledge and practice.

Methods The simulator was designed to replicate an outpatient psychiatric clinic experience. The virtual patient reported symptoms of a treatment-resistant form of major depressive disorder (MDD), requiring the learner to use various antidepressants in order for the patient to fully remit. Test scores were based on the proportion of correct responses to questions asked by the virtual patient about possible side effects, dosing, and titration decisions, which depended upon the patient’s tolerability and response to the learner’s selected medications. The validation paradigm included a novice-expert performance comparison across 4th year medical students, psychiatric residents from all four post-graduate year classes, and psychiatry department faculty, and a correlational analysis of simulator performance with the PRITE Somatic Treatments subscale score. Post-test surveys evaluated the test takers’ subjective impressions of the simulator.

Results Forty-three subjects completed the online exam and survey. Total mean scores on the exam differed significantly across all the learner groups in a step-wise manner from students to faculty ($F = 6.10, p = 0.0001$). Total mean scores by residency class correlated with PRITE Somatic Therapies subscale scores ($p < 0.01$). The post-test survey mean Likert results ranged from 3.33 ± 1.20 to 4.4 ± 0.79 , indicating neutral to favorable responses for use of the simulator.

Conclusions This simulator demonstrated strong construct validity and high participant acceptability for assessing proficiency in the psychopharmacologic treatment of MDD.

Keywords Residents · Psychopharmacology · Simulation · Assessment

The last three decades of outpatient psychiatry practice have witnessed an increasing emphasis on psychopharmacology over psychotherapy [1, 2]. Simultaneously, the armamentarium of psychotropic medications has grown substantially, with over 20 medications now having marketing approval for the treatment of major depressive disorder alone. Illnesses previously believed to be best treated primarily by psychotherapy, such as substance use disorders [3] and some types of eating disorders [4], now have pharmacological treatment options. Moreover, the continuing expansion of mechanisms of action [5, 6] utilized by newly approved drugs further increases the

psychopharmacology knowledge requirements for psychiatric drug prescribers.

Theoretical education in psychopharmacology in psychiatry residency programs is delivered through didactic lectures and journal clubs, while practice-based learning occurs through inpatient and outpatient psychiatry experiences which provide opportunities to initiate medicines, monitor treatment responses, and manage emerging side effects [7]. Currently, residents’ psychopharmacology proficiency is measured by their performance on the Psychiatry Residency In Training Exam (PRITE) and through the Accreditation Council for Graduate Medical Education (ACGME) Psychiatry Milestone assessments completed by supervisory attendings. However, these tools have significant limitations regarding their ability to assess trainees’ psychopharmacologic knowledge. The PRITE is a 300 question exam that psychiatric residents take annually [8]. It assesses knowledge in a variety of areas of psychiatry, including diagnostic assessment,

✉ Jeffrey J. Rakofsky
jrakofs@emory.edu

¹ Emory University, Atlanta, GA, USA

² University of Southern California, Los Angeles, CA, USA

epidemiology, and neurology, but psychopharmacology constitutes < 15% of the questions. Additionally, the questions are primarily presented using a short patient vignette followed by a single-question, multiple-choice format. The vignettes do not provide follow-up information or additional questions after the initial question is answered. Consequently, the PRITE structure provides for only very limited assessment of the depths of residents' psychopharmacology knowledge and ability to use medication in a manner that reflects real-world practice.

The ACGME Psychiatry Milestones is an assessment tool that provides a framework for evaluating behaviors or qualities associated with a resident's development as a physician [9]. Attending physicians who work with residents complete the form, indicating the level of knowledge or practice skills the resident has achieved during the period of assessment. For psychopharmacologic knowledge, this tool includes a single item, "PC5 Somatic Therapies," which subsumes a number of behaviors related to "using psychopharmacologic agents in treatment." The scoring range for the item is 1–4, which lacks adequate scale to capture the complexity of the psychopharmacologic practice. For example, behaviors listed as part of the highest level of competency, level 4, such as titrating dosages and managing side effects, are often demonstrated by first year residents working on inpatient units. As a result, this milestone has very limited ability to discriminate junior from senior residents and to identify specific areas of deficiency. A second Milestones item, "MK5. Somatic Therapies," requires faculty to rate the resident's medical knowledge of somatic therapies, including medicines, electroconvulsive therapy, and other emerging somatic therapies. Ratings on this item are impressionistic rather than systematic, given the large number of somatic therapies that exist and the reliance on clinical discussions between resident and attending to make this determination. The Milestones tool is also susceptible to the recency effect [10], given that the faculty member's assessment is retrospective and thus likely to overweigh recent or highly salient events or interactions. These limitations of existing assessments point to the need for more specific, in-depth, content-valid assessment of residents' psychopharmacology knowledge and skills.

An alternative to paper exams and faculty impressions is assessments conducted via simulators. When used for assessment purposes, simulators have the advantage of eliminating assessor biases because scoring is systematic and based on the presence or absence of specific actions. Additionally, high-fidelity simulators may provide realistic testing scenarios that can more fully assess skills and knowledge than a multiple-choice question-based exam. To date, simulators developed for psychiatric uses have focused on enhancing or assessing students' communication and diagnostic skills; none, as far as the authors know, have been developed to specifically assess psychiatry residents' proficiency in medication management for depression [11].

The goal of this study was to develop an evaluative tool that could eventually replace the PRITE and other forms of theoretical evaluation. Herein, we report on a virtual standardized patient (VSP)-based psychopharmacology simulator developed to provide a summative assessment of the learner's ability to initiate medication, adjust doses, and manage the emerging side effects in a patient with treatment-resistant major depressive disorder.

Methods

Exam development requires collecting validity evidence to evaluate the appropriateness of the use, interpretations, or decisions that arise from the exam results [12]. The Kane framework for testing validity arguments for educational assessments organizes the evidence into four categories: *scoring*, *generalization*, *extrapolation*, and *implications* [13]. Scoring pertains to how the test performance is translated into a score, generalization pertains to how the score reflects test performance, extrapolation pertains to how the score reflects real-world performance, and implications pertain to how the score influences decisions or actions that affect the learner (e.g., promotion, remediation). This study focused primarily on collecting extrapolation data in the form of novice-expert performance comparisons and comparison to a standardized test measuring a similar construct. These comparisons are the most common approach when validating the use of medical simulators [12]. The Emory University Institution Review Board designated this study to be exempt from review.

Software Development

The virtual standardized patient software was created through the University of Southern California Standard Patient Studio platform, a freeware virtual patient community that was developed by the University of Southern California with funding from the Department of Defense. Standard Patient combines virtual human avatars, artificial intelligence, and an advanced pedagogical design to create realistic, emotionally expressive interactions, including live voice communication. In addition to providing conversational interaction, the system supported live feedback to subjects and collected a myriad of performance parameters. Prior research has demonstrated that Standard Patient has shown a high degree of performance, assessment accuracy, and utility for training [14, 15]. See Fig. 1.

Content Development

The patient narrative was created by one of the authors (JJR), who is a mood disorders expert and educator, with more than 15 years of clinical experience, more than 8 years of directing

and providing supervision in a residents' psychopharmacology clinic, and with more than 40 peer-reviewed publications on medical education, major depression, and bipolar disorder. The narrative featured a 28-year-old white man with major depressive disorder (MDD) who sees a psychiatrist for medication management. The story line was divided into four sequential modules, with each module featuring particular classes of medications. The four modules were generally aligned with the treatment algorithm applied in the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study [16], while also incorporating newer pharmacologic options based on more recent studies. The first module included selective serotonin reuptake inhibitors and bupropion. The second module included serotonin-norepinephrine reuptake inhibitors and mirtazapine. Augmentation strategies (e.g., lithium, triiodothyronine, and second-generation antipsychotics), tricyclics, and monoamine oxidase inhibitors were included in the third and fourth modules. Within each module were "distractor" psychotropic medication options that most psychiatrists would not prescribe for MDD at that level of treatment resistance. The learner was prompted to select a medication in each module to target the patient's symptoms. Following the selection, the virtual standardized patient would ask the learner questions about the medicine, including possible side effects and dose titration questions. Food and Drug Administration medication package inserts and clinical trials data were used to generate the correct answers to these questions. The learner would select from a list of possible answer

choices and was given immediate feedback on all of their choices and the rationale. If the virtual patient agreed to take the medicine, he would return 1 month later (instantly for the test taker) to provide an update on the effect and tolerability of the medicine. This process continued until the learner exhausted the treatment options for the particular module and then was instructed to move on to the following module. Because the virtual patient's depression was designed to be treatment-resistant, the learner was forced to move through all 4 modules of the software.

Throughout the different modules, the learner was asked a variety of questions about the medicines they selected. These questions pertained to starting doses, possible side effects, relevant lab work, and dose titration decisions in the face of non-response or tolerability problems. At the least, there were 16 questions each focusing on dosing, side effects, and dose titration decisions. The more medicines selected by the participant, the more questions they were asked.

A pilot version of the exam was taken by an expert (BWD) in the psychopharmacology of mood disorders who has more than 20 years of clinical experience, more than 8 years of supervising residents in a psychopharmacology clinic, and over 60 publications on the biology and treatment of MDD. The pilot exam was also taken by a chief resident in psychiatry and 4th year medical students completing a digital medicine elective offered by the University of Southern California. They were instructed to provide feedback on the breadth of medicines included, the clarity of the virtual patient's

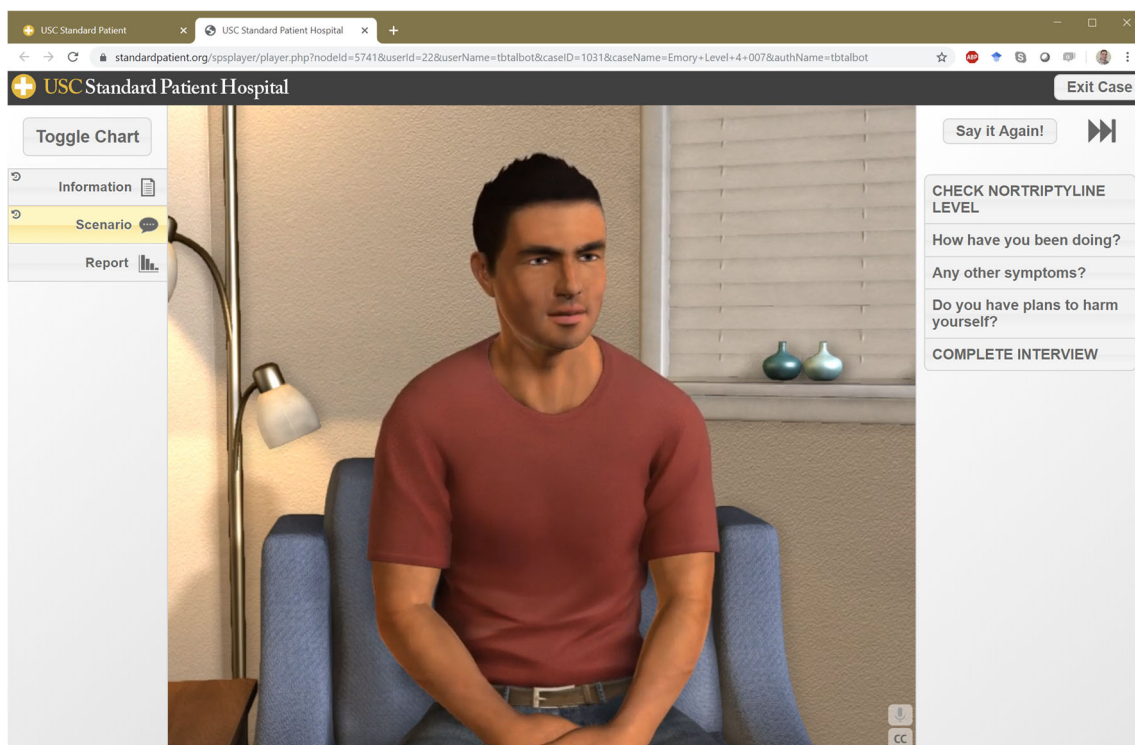


Fig. 1 A screen shot image of the virtual standardized patient

communication, and to look for any errors in the questions, response options, or general flow of the software. Their feedback was collected and used to make revisions to the simulator.

Recruitment and Testing

Emory University fourth year medical students (M4) participating in a psychiatry sub-internship elective, general psychiatry residents (PGY level 1–4), and faculty with psychopharmacology practices were recruited via institutional e-mail to participate in the testing. They were offered a \$25 gift card to reimburse them for their time. Participants were informed that their test results were anonymous and could not be linked to them individually, so their answers could have no impact on their academic standing. Learners took the exam without access to supplemental materials.

Multiple supervised test sessions were scheduled to occur from January–February 2019. These sessions were held in a classroom within the residency education suite, and participants could attend the session that was most convenient for them. Participants were required to bring a laptop and ear buds. At the beginning of the test session, a proctor played a pre-recorded online video that reviewed the instructions for completing the exam. The purpose of the video was to ensure consistency of instruction across study sessions. Use of the closed-caption feature on the simulator was encouraged but not required. To avoid contamination, participants were instructed to avoid discussion of the test with others after completing the exam and to refrain from use of their smartphones or the internet for assistance while taking the exam. Participants were then provided a handout which listed their randomly generated username, password, the weblink to the exam, and the weblink to the post-test survey. The proctor remained in the room for the duration of the exam to resolve technical problems and to hand out the gift card upon completion of the exam and post-test survey. Participants were given up to 60 min to complete the exam although everyone finished within 30–60 min.

Given the interest in participating but the inconvenience of the testing sessions for some, towards the end of the recruitment period, we permitted participants to take the exam in an unsupervised setting, at a time and location of their choosing. The protocol was the same as it was with the supervised testing sessions except no proctor was in the room while the participants completed the exam.

Correlation with a Test Measuring a Similar Construct

The PRITE exam scoring report provided to residency program directors breaks down each resident's score into three different categories: Global Scores, Psychiatry Subscale Scores, and Milestones. A number of subscale scores are

reported within each of these categories. In the Milestones category, the MK5: Somatic Therapies subscale most closely reflects the construct being measured by the virtual patient simulator. Grouping the results by class and deidentifying the resident, the Emory University residency program provided the 2018 MK5: Somatic Therapies subscale standardized scores of all the psychiatry residents who participated in the virtual standardized patient assessment. Because residents' simulator performance was identified only by their residency class, the correlational analyses had to be conducted by residency class mean performance rather than by individual resident performance. Faculty and medical students were excluded from this analysis since they did not take the PRITE exam.

The 2018 PRITE exam included 32 MK5: Somatic Therapies questions out of 300 total questions. Of those 32, nine (28%) referenced an antidepressant or non-bipolar depression in the question stem, and another five (16%) included at least one antidepressant as a possible answer response option.

Post-Test Survey

A ten-item survey to assess test acceptability to the learner was created. For the first eight questions, respondents were asked to use a Likert scale (1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, 5 = strongly agree) to indicate the degree to which they agreed or disagreed with statements about the simulator. The ninth question pertaining to the pacing of the test provided respondents with answer choices "too fast," "too slow," or "just right," and the tenth question was an open-ended inquiry for written responses to their experience of taking the test. No identifying information was solicited on the survey in order to maximize the respondent's candor.

Statistical Methods

Each question on the exam assessing psychopharmacology knowledge and decision-making was worth one point, and scores were calculated by dividing the number of points by the total number of questions on the exam. This total number was a function of the number of medicines each participant selected in each of the four modules. Participants with incomplete data on any modules of the exam were excluded from the analysis. Score means and standard deviations were calculated for each participant group (M4, PGY-1, PGY-2, PGY-3, PGY-4, faculty) for each module of the exam, and for the total performance on the exam (all modules combined). ANOVA testing was used to compare the overall scores between each participant group followed by a post hoc Tukey's test to determine which groups were significantly different. Percentage of questions answered correctly for items related to side effects, dosing, and titration were also reported for each participant group. Means for each residency class were calculated for the PRITE

MK5: Somatic Therapies subscale standardized scores and then correlated with the mean total score on the simulator for each residency class using Spearman's correlation coefficient. Likert scores for each applicable item on the post-test survey were averaged and reported. The data were analyzed using SPSS Statistics 26 (IBM, Armonk, New York).

Results

Sample Description

Forty-three subjects completed the online exam and survey with 81% (30 of 37) of the training program's residents participating; exam data from three subjects were excluded due to incomplete responses. The analyzed sample included 6 fourth-year medical students (15% of the total sample), 5 PGY-1 residents (12.5%), 10 PGY-2 residents (25%), 6 PGY-3 residents (15%), 8 PGY-4 residents (20%), and 5 faculty members (12.5%). The majority of participants were male ($n = 25$, 62.5%; female $n = 15$, 37.5%) and completed the exam and survey in a supervised testing setting ($n = 27$, 67.5%; unsupervised, $n = 13$, 32.5%). The 13 participants who completed the exam unsupervised included 6 faculty (only 5 with complete data), 3 PGY-3s, and 4 medical students.

Exam Results

As shown in Fig. 2, total mean scores and standard deviation on the exam differed significantly among the learner groups in a step-wise manner: M4 student = 69.7 ± 7.6 , PGY-1 = 69.2 ± 6.2 , PGY-2 = 79.5 ± 8.3 , PGY-3 = 82.4 ± 5.9 , PGY-4 = 83.0 ± 6.9 , faculty = 86.8 ± 5.0 ($F = 6.1$, $p < 0.001$). Post hoc testing revealed significant mean score differences among the following groups: PGY-3 vs. M4 (12.7 ± 4.1 , $p = 0.03$), PGY-3 vs. PGY-1 (13.2 ± 4.3 , $p = 0.04$), PGY-4 vs. M4 (13.3 ± 3.8 , $p =$

0.02), PGY-4 vs. PGY-1 (13.8 ± 4.0 , $p = 0.02$), Faculty vs. M4 (17.1 ± 4.3 , $p = 0.004$), and Faculty vs. PGY-1 (17.6 ± 4.4 , $p = 0.005$). Although faculty scored higher than senior-level residents, the differences were not statistically significant, suggesting that the senior residents had achieved an acceptably high level of psychopharmacologic knowledge.

The percentages of side effect, dosing, and dose titration questions answered correctly are presented in Table 1. There was a linear trend for greater accuracy within increasing levels of learner experience for all three types of psychopharmacology questions.

Correlation Results

PRITE MK5: Somatic Therapies subscale scores were received for 27 of the 30 participating residents. The mean score per class and standard deviation was PGY-1 = 419.4 ± 85.8 , PGY-2 = 490.6 ± 54.0 , PGY-3 = 572.5 ± 69.8 , and PGY-4 = 617.7 ± 97.2 and correlated significantly with the mean total simulator performance by class ($p < 0.01$). See Fig. 3.

Survey Data

As shown in Table 2, the post-test survey mean results ranged from 3.3 (1.2) to 4.4 (0.8), indicating overall favorable responses for most components of the simulator, and a neutral response when compared to an oral exam. None of the learners thought the test moved too quickly. Sixty percent (25/42) of participants described the pacing of the test as "just right," while 40% (17/42) described it as "too slow."

Discussion

This study evaluated a novel, virtual standardized patient-based psychopharmacology assessment simulator among

Fig. 2 Comparison of mean total score performance among all participant groups. key = M4 ($n = 6$), PGY-1 ($n = 5$), PGY-2 ($n = 10$), PGY-3 ($n = 6$), PGY-4 ($n = 8$), faculty ($n = 5$), ANOVA results, $F = 6.1$, $p = 0.0001$

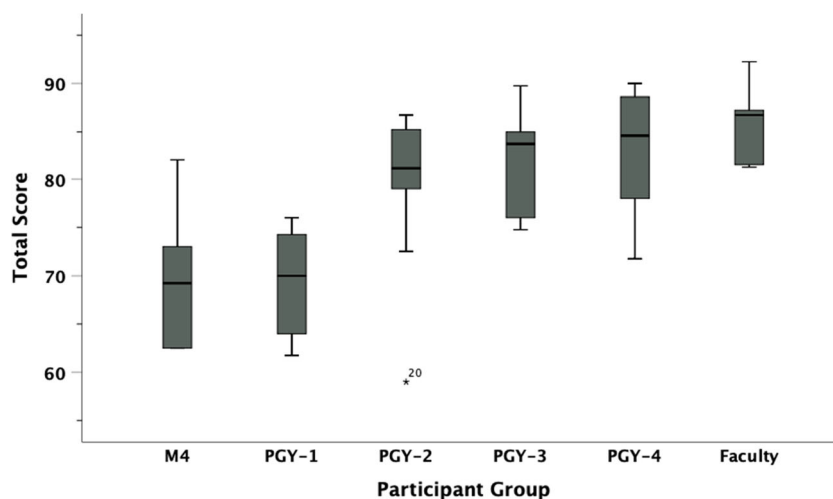


Table 1 Percentage of questions answered correctly within psychopharmacology areas of knowledge among the different participant groups

Areas of knowledge:	Participant group:					
	M4	PGY-1	PGY-2	PGY-3	PGY-4	Faculty
Side effects	71%	74%	83%	81%	94%	98%
Dosing	44%	49%	69%	77%	67%	69%
Dose titration	70%	70%	75%	79%	73%	83%

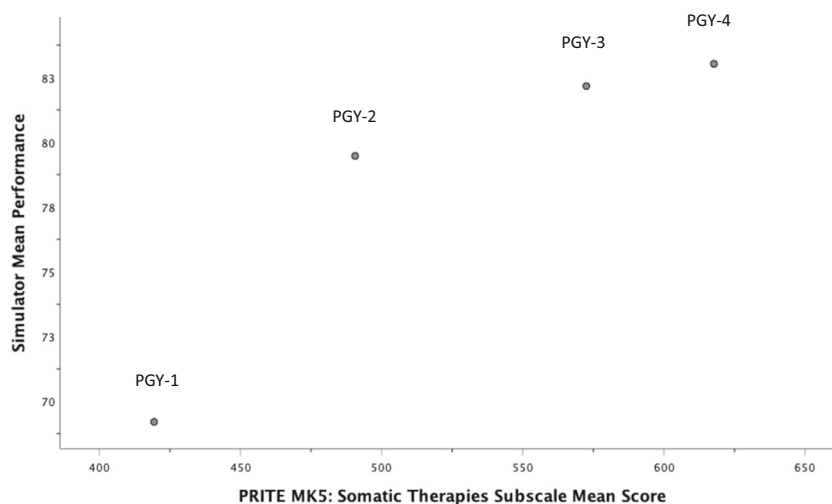
Areas of knowledge are the categories of psychopharmacology knowledge that clinicians must have to prescribe psychotropics safely and effectively. The participant groups include the trainees (medical students and residents) and faculty members who completed the virtual standardized patient assessment

medical students, psychiatry residents, and psychiatry attendings within an academic medical center setting. The results demonstrated that the simulator had strong construct validity using novice-expert performance comparisons and comparisons to a test measuring a similar construct and had high participant acceptability. The participant groups with greater experience prescribing psychotropic medications performed better than those with the least experience. Faculty, PGY-4, and PGY-3 residents all performed statistically significantly better than the PGY-1 residents and 4th year medical students; however, there were numeric differences between all groups with faculty scoring the highest. The step-wise improvement in scores by experience level suggests that this tool can discriminate between test takers with different levels of mastery, supporting the tool's construct validity as a measurement of proficiency in the psychopharmacological treatment of major depressive disorder. This step-wise improvement between groups was also seen when focusing on the specific areas of psychopharmacology knowledge: side effects and titration. Medication dosing showed some step-wise improvement; however, the PGY-3 residents excelled in this category, likely owing to their recent experiences working in a psychopharmacology clinic with a high volume of mood disorder patients.

The mean total score on the simulator by class correlated with the mean scores on the Somatic Therapies subscale of the PRITE, providing additional construct validity evidence for the virtual-patient simulator. Although the PRITE is not a comprehensive measure of psychopharmacology knowledge, the Somatic Therapies subscale focuses on psychopharmacology. Additionally, the exam itself is a widely used tool to measure overall resident competence and correlates with future performance on the American Board of Psychiatry and Neurology Part I exams [17].

On the user experience survey, average scores for the survey items fell between the “neutral” and “strongly agree” anchor points. The lowest rated (neutral) item was for the statement, “this test is a superior way to assess clinical psychopharmacology skills compared to an oral exam”. The survey comments suggest that some participants believed an oral exam would give them more latitude with answer responses as they could justify their choices. While this is true, the disadvantages to using oral exams include more opportunity for examiner bias, less standardization, and more personnel requirements. The highest rated item was the statement, “the process of moving through the test was intuitive and clear”, supporting the usability of this tool and its ability to reliably engage test takers. The item, “the experience treating this

Fig. 3 Correlation of psychiatry residency class mean performance on the virtual standardized patient simulator with mean scores on the PRITE MK5: somatic therapies subscale. Spearman correlation coefficient = 1, $p < 0.01$, PGY = post graduate year



virtual standardized patient is reasonably similar to the outpatient psychiatry experience” was rated in the neutral to agree range, supporting the authenticity of the simulator. Given the number of participants who likely had minimal to no outpatient psychiatric experiences up to that point (e.g., fourth year students, PGY-1 and PGY-2 residents), it is possible that this item may have been scored even more favorably had it been limited to the senior level residents and faculty. Because the surveys were anonymous, this possibility could not be explored. Examples of constructive survey comments included, “It would be nice to have a wider array of interview questions to choose from for the patient interview component,” and “It would’ve been enhanced by the test taker being able to offer free text or other alternatives.”

A potential limitation to this study was the inclusion of only those fourth year medical students participating in psychiatry sub-internship electives. Had a broader group of fourth year medical students been recruited, it is possible there would have been a greater separation in the scores between the medical students and other participant groups. Another limitation was the small sample sizes per participant group, which limited power to identify statistically significant differences between each experience level. Because we decided to ensure participants’ anonymity to maximize candor, and due to the lack of a gold standard to measure psychopharmacology proficiency, we could not compare test performance to learners’ real-world outcomes.

Strengths of the study include the method and simulator design features that address the different validity components within the Kane Framework for testing validity arguments. Those features include (1) *Scoring*: computer-based entry of answers to accurately capture and score learner performance; the use of equally weighted scoring for each question to reduce bias among the different areas of psychopharmacology knowledge; survey data indicating that the test-taking process and the virtual patient’s voice was clear; pilot phase testing to determine usability; and the use of a proctor to prevent cheating and test contamination; (2) *Generalization*: test

question development generated from a standard treatment algorithm; questions covering the major components of psychopharmacology (dosing, side effects, titration); and pilot testing to determine adequate breadth of the questions; (3) *Extrapolation*: the novice-expert performance comparisons; correlation with a test measuring a similar construct; survey data indicating that the simulator experience was similar to the outpatient experience; the use of a realistic-appearing outpatient psychiatric office in the design of the software and the use of natural prosody in the virtual patient’s voice to enhance the authenticity of the testing experience. Addressing the *Implication* of the simulator test results will require evaluation as a formal assessment tool within a residency program.

The development and validation of this psychopharmacology summative assessment tool demonstrates the potential utility of a virtual standardized patient simulator to achieve a fair and full evaluation of residents’ psychopharmacology proficiency in treating MDD. The major advantage to this kind of exam is its ability to reduce bias from the assessor and to evaluate psychopharmacology knowledge in an in-depth, realistic, and dynamic way. Because a computer-based simulator can scale up for wider use easily across institutions, it is conceivable that with more validity data collected over time and in larger samples, a limited suite of similar simulators, including those developed for other psychiatric disorders, could in combination yield a standardized summative assessment of psychopharmacology proficiency across residencies. This assessment could occur throughout the 4th year of residency as senior residents would have had a substantial number of hospital and clinic training opportunities to prepare them for such a comprehensive exam. For this vision to be achieved, virtual standardized patient simulators testing proficiency in the treatment of other psychiatric illnesses (e.g., schizophrenia, bipolar disorder) will need to be developed, and more validation studies addressing all the components of Kane’s framework will be required to support the use of these simulators as a standard measure of psychopharmacology knowledge and skills. Finally, in the age of the COVID-

Table 2 Post-exam survey results using a Likert scale (1 = strongly disagree, 5 = strongly agree)

Survey item	Likert score
1. I felt the test accurately assessed my knowledge of the pharmacologic treatment of major depression	4.1
2. This test is a superior way to assess clinical psychopharmacology skills compared to a pen and paper test	4.0
3. This test is a superior way to assess clinical psychopharmacology skills compared to an oral exam	3.3
4. The experience of treating this virtual patient is reasonably similar to the outpatient psychiatry experience	3.4
5. This test involves medications typically used for the treatment of Major Depressive Disorder	4.4
6. The process of moving through the test was intuitive and clear	4.3
7. The speech of the virtual patient was sufficiently clear	3.8
8. The test was not unnecessarily repetitive	3.4

The survey items are the statements included in the post-test survey that respondents completed after taking the virtual standardized patient assessment. The Likert score reflects respondents’ mean level of agreement with the survey statements

19 pandemic, this virtual online assessment tool and others like it can allow evaluations of residents to occur remotely without creating an increased risk of infection for residents, patients, and faculty.

Funding Information Funding for this study was provided by the American Board of Psychiatry and Neurology “Faculty Innovation in Education Grant” awarded to JJR.

Compliance with Ethical Standards

Disclosures JJR receives research funding from Takeda and National Institutes of Mental Health.

TBT owns Medical Mechanica LLC which licenses the software used for this study.

BWD receives research support from Acadia, Sage, Takeda, and the National Institutes of Health, and has served as a consultant to Myriad Neuroscience and Aptinyx, none of which have any commercial interest in this software.

References

- Mojtabai R, Olfson M. National trends in psychotherapy by office-based psychiatrists. *Arch Gen Psychiatry*. 2008;65:962–70.
- Olfson M, Marcus SC, Wan GJ, et al. National trends in the outpatient treatment of anxiety disorders. *J Clin Psychiatry*. 2004;65:1166–73.
- Lee JD, Nunes EV Jr, Novo P, et al. Comparative effectiveness of extended-release naltrexone versus buprenorphine-naloxone for opioid relapse prevention (X:BOT): a multicentre, open-label, randomized controlled trial. *Lancet*. 2018;391:309–18.
- Hudson JI, McElroy SL, Ferreira-Cornwell MC, et al. Efficacy of lisdexamfetamine in adults with moderate to severe binge-eating disorder: a randomized clinical trial. *JAMA Psychiat*. 2017;74:903–10.
- Popova V, Daly EJ, Trivedi M, et al. Efficacy and safety of flexibly dosed esketamine nasal spray combined with a newly initiated oral antidepressant in treatment-resistant depression: a randomized double-blind active-controlled study. *Am J Psychiatry*. 2019;176:428–38.
- Gunduz-Bruce H, Silber C, Kaul I, et al. Trial of SAGE-217 in patients with major depressive disorder. *N Engl J Med*. 2019;381:903–11.
- Rakofsky JJ, Dunlop BW. US psychiatric residents’ treatment of patients with bipolar disorder. *J Clin Psychopharmacol*. 2012;32:231–6.
- The American College of Psychiatrists Available from: <https://www.acpsych.org/prite>. Accessed 15 Jan 2020.
- The Psychiatry Milestone Project: a joint initiative of the American Council for Graduate Medical Education and the American Board of Psychiatry and Neurology. 2015. <https://www.acgme.org/Portals/0/PDFs/Milestones/PsychiatryMilestones.pdf>. Accessed 15 Jan 2020.
- Murdock B. The SERIAL POSITION EFFECT OF FREE RECALL. *J Exp Psychol*. 1962;64:482–8.
- Williams B, Reddy P, Marshall S, et al. Simulation and mental health outcomes: a scoping review. *Adv Simul*. 2017;2:2.
- Cook DA, Hatala R. Validation of educational assessments: a primer for simulation and beyond. *Adv Simul*. 2016;1:31.
- Cook DA, Brydges R, Ginsburg S, et al. A contemporary approach to validity arguments: a practical guide to Kane’s framework. *Med Educ*. 2015;49:560–75.
- Talbot T, Rizzo A. Virtual standardized patients for interaction conversational training: a grand experiment & new approach. *Exploring the Cognitive, Social, Cultural, and Psychological Aspects of Gaming and Simulations*: IGI Global; 2019.
- Talbot T. USC Standard Patient Research Report. 2015. Accessed at <https://www.dropbox.com/s/wjn4s6teh4bzja2/2015%20USC%20Standard%20Patient%20Research%20Report.pdf?dl=0>. Accessed 15 Jan 2020.
- Gaynes BN, Rush AJ, Trivedi MH, et al. The STAR*D study: treating depression in the real world. *Cleve Clin J Med*. 2008;75:57–66.
- Webb LC, Juul D, Reynolds CF 3rd, et al. How well does the psychiatry residency in-training examination predict performance on the American Board of Psychiatry and Neurology. Part I. Examination? *Am J Psychiatry*. 1996;153:831–2.

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.