Evaluation of a virtual agent to train medical students conducting psychiatric interviews for diagnosing major depressive disorders

Lucile Dupuy[1], Jean-Arthur Micoulaud-Franchi[1,2], Hélène Cassoudesalle[2], Orlane Ballot[3], Patrick Dehail[1,2], Bruno Aouizerate [1,2], Emmanuel Cuny [1,2], Etienne de Sevin[1], Pierre Philip[1,2]

[1] University of Bordeaux, USR 3413 SANPSY, 33076 Bordeaux Cedex, France
[2] University Hospital of Bordeaux, CHU Pellegrin - Tripode, 33076 Bordeaux Cedex, France
[3] University of Laval, Centre d'étude des troubles du sommeil, Québec, G1V 0A6, Canada

Corresponding author:
Lucile Dupuy, Ph.D
USR CNRS 3413 SANPSY– Sommeil, Addiction et Neuropsychiatrie
University of Bordeaux
Site Carreire – Zone Nord, Bat 3B, 3rd floor
33076 Bordeaux Cedex, France
Email: lucile.dupuy@u-bordeaux.fr
Phone number: +33 5 57 57 11 00

Evaluation of a virtual agent to train medical students conducting psychiatric interviews for diagnosing major depressive disorders

Abstract

Background: A psychiatric diagnosis involves the physician's ability to create an empathic interaction with the patient in order to accurately extract semiology (i.e., clinical manifestations). Virtual patients (VPs) can be used to train these skills but need to be evaluated in terms of accuracy, and to be perceived positively by users.

Methods: We recruited 35 medical students who interacted in a 35-min psychiatric interview with a VP simulating major depressive disorders. Semiology extraction, verbal and non-verbal empathy were measured objectively during the interaction. The students were then debriefed to collect their experience with the VP.

Results: The VP was able to simulate the conduction of a psychiatric interview realistically, and was effective to discriminate students depending on their psychiatric knowledge. Results suggest that students managed to keep an emotional distance during the interview and show the added value of emotion recognition software to measure empathy in psychiatry training. Students provided positive feedback regarding pedagogic usefulness, realism and enjoyment in the interaction.

Limitations: Our sample was relatively small. As a first prototype, the measures taken by the VP would need improvement (subtler empathic questions, levels of difficulty). The face-tracking technique might induce errors in detecting non-verbal empathy.

Conclusion: This study is the first to simulate a realistic psychiatric interview and to measure both skills needed by future psychiatrists: semiology extraction and empathic communication. Results provide evidence that VPs are acceptable by medical students, and highlight their relevance to complement existing training and evaluation tools in the field of affective disorders.

Keywords

Virtual Patient; Medical training; User experience, Psychiatric Interview; Emotion Detection

1. Introduction

Clinical diagnosis relies on both *signs* (i.e., externally, observable phenomena - expressions) and *symptoms* (i.e., patient's subjective complaints - experiences) (Nordgaard et al., 2013) in order to perform semiology[1] extraction (i.e., "*clinical evaluation of signs and symptoms, leading to the identification of a disorder*") (Micoulaud-Franchi et al., 2018). In most medical specialties, physicians use tools to measure signs (e.g., blood pressure monitor, medical imaging), and clinical interviews to collect subjective symptoms. In the field of affective disorders, however, the vast majority of signs investigated, such as body movements, language and discourse, are expressed as patients progressively disclose their symptoms, so the psychiatrist has to rely on his/her ability to conduct an appropriate clinical interview in order to disentangle and collect both the signs and symptoms of their patients (Shea, 2016; Silverman et al., 2015). In this context, the psychiatric interview should be conversational, contextually adapted and empathic (Nordgaard et al., 2013). Notably, Shea (Shea, 2016) suggests that a first psychiatric interview should follow three main phases:

1. *The introduction and beginning of the interview*, which aims to lower the patient's anxiety of coming to see a psychiatrist, and expose the objectives of the present interview;
2. The *main part of the interview*, during which objectives are to help the patient express his symptoms by guiding him through the different dimensions of depressive disorders;
3. The *ending of the interview*, where the psychiatrist presents the diagnosed disorder and proposes an adapted solution, while taking into account the patient's feelings and representations and giving him hope about future recovery.

Therefore, two major skills have to be acquired by future psychiatrists. First, the ability to extract semiology based on their knowledge of clinical signs and symptoms for each category of mental and affective disorder as listed in the DSM-5 (American Psychiatric Association, 2013). However, as importantly, the psychiatrist needs to create an empathic relationship with the patient (Bhugra et al., 2017; Plakun, 2015) in order to facilitate the procedure. Empathy is arguably the most important psychosocial characteristic of a physician engaged in patient care (Colliver et al., 2010), as it helps build patient trust (Deladisma et al., 2007), increases patient satisfaction and compliance, improves medical care outcomes and may reduce medical malpractice lawsuits (Kim et al., 2004). Based on the literature,

[1] The term semiology is used in this article as a synonym of the term symptomatology, presentation, manifestation or phenomenology of the disorder. In this article, we use the term semiology, in line with the French medical tradition from the early 19th century. In the English tradition, the term semiology has referred since the middle of the 17th century to the science of language.

we distinguish two types of empathy: *verbal empathy,* referring to the physician's ability to 'help the patient express his/her symptoms' (Shea, 2016); and *nonverbal empathy*, corresponding to the physician's ability to stay neutral (Nordgaard et al., 2013) and to show empathic listening (Plakun, 2015).

Medical education consists primarily in passive learning through lecture-based classroom and clinical observation, which has demonstrated poor performances in remembering (Tolks et al., 2016). Complementarily, new techniques are now being used to improve students' empathic skills (for a review, see Batt-Rawden et al., 2013), mainly provided by role play with standardized patients (SPs), *i.e.* actors trained to act as patients. However, even if these initiatives have been effective in improving medical students' empathy, they are sometimes not feasible in terms of schedule and resources to train and employ. Additionally, assessment methods need to evaluate students' abilities to conduct an empathic interview with a patient. New tools are therefore needed to provide future psychiatrists with active, practical and experiential training and assessment while remaining feasible with time and resource constraints, standardized, and common to all medical schools (Bhugra et al., 2017).

In this context, computerized tools are regarded as a promising solution to provide new tools for training and assessment in medical education. Notably, embodied conversational agents (ECAs), defined as *"virtual digital representations of a computer interface in the form of human-like faces"*(Cassell et al., 2000), are now being developed for use as virtual patients (VPs) in medical training (Cook et al., 2010). Notably, in their recent study, Maicher's team (Maicher et al., 2019) developed a VP to train medical students' information-gathering skills. Results with 102 students showed that the VP was comparable to human raters to evaluate information-gathering skills. However, the interaction was based on typed text, thereby precluding all forms of nonverbal and empathic interaction. In a randomized controlled trial with 70 first-year medical students, Foster and colleagues (Foster et al., 2016) found that students interacting through a text-based interface with a depressive VP giving feedback about empathic responses were later able to be more empathetic in an interaction with an SP. Another study (Kleinsmith et al., 2015) found that students were more empathetic to VPs than to SPs. They considered that they were being judged less, felt less stressed that they were not dealing with real patients, and had more time to think about their answer. Similarly, in (Deladisma et al., 2007), students felt less nervous than when talking with a real SP after interacting with a life-size VP suffering from abdominal pain. Taken together, these findings show that VPs can provide problem-oriented, standardized, repetitive and safe practice that simulate cases not possible for human actors (*e.g.,* facial paralysis), while providing situations less stressful for students and with no consequences for patients. However, until now, only a few VPs have been developed and tested, they simulate only short and non-realistic (mostly text-based) interviews, and none of them has focused on both semiology extraction and the assessment of empathy during the interview. Therefore, their applicability to training for conducting psychiatric interviews is limited. Despite the high prevalence of depression (Bromet et al., 2011), only one study focused on a VP suffering from this disorder. Moreover, depression symptomatology exhibits both cognitive and motor dimensions (Kaplan and Sadock, 1988), suggesting its appropriateness for simulation. For these reasons, the objectives of this study were to design and validate a realistic psychiatric interview with a VP simulating major depressive disorders, and to assess medical students' skills in conducting an interview, in terms of semiology extraction of depression and empathic communication.

2. Methods and Materials

2.1 Participants

Thirty-five students were recruited from June 2016 to July 2017. They all were fourth-year medical students[2] from Bordeaux Medicine school (France), were 22 years of age on average, and half of them (N = 17) were male. They were recruited during their obligatory fourth-year observational clinical training course in Bordeaux University Hospital. Among them, 15 were trainees in the psychiatry department (and thus had already observed psychiatric interviews) and 20 in the neurology department (therefore never having experienced a psychiatric interview).

This project is part of a larger project on virtual reality and clinical phenotyping (PHENOVIRT) that has been approved in compliance with French and European regulations on clinical research by a local ethics committee (*Comité pour la Protection des Personnes* – Institutional Review Board of University of Bordeaux). All participants gave their written informed consent before entering the study.

2.2 The Virtual Patient

Our Virtual Patient (VP) portrays a middle-aged woman suffering from major depressive disorder (MDD) according to the DSM-5 criteria (American Psychiatric Association, 2013) (**Figure 1**). The implementation of the VP and the interaction scenario were created to give a realistic example of depressive symptoms in order to promote sensorial, emotional and episodic memorization. We placed

[2] In France, medical school starts directly after high school, thus fourth-year French medical students, often called « externs » or "hospital students", correspond to first-year medical students in the US curriculum. During the fourth year, "externs" generally have to spend five mornings per week in several specialty departments, under the responsibility of a senior physician, to learn how to recognize the various signs of a disease.

special emphasis on the prosody, gestures, and general aspect of the VP, by involving an actress (who was psychologist as well and had experience with depressive patients) and to capture her voice and her non-verbal behavior with motion capture technology. In order to provide ecological conditions, the VP was displayed on a TV screen in the size of a real human (**Figure 1**). To record students' verbal and non-verbal behavior during the interview, their face was recorded and analyzed by emotion recognition software (see 2.4.1.3).

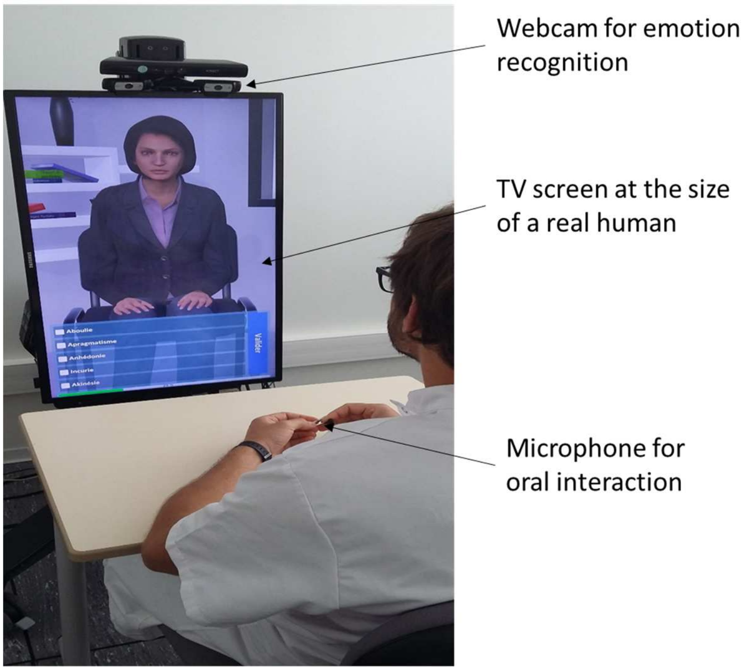

Webcam for emotion recognition
TV screen at the size of a real human
Microphone for oral interaction
Aboulie
Apragmatisme
Anhédonie
Incurie
Akinésie
Valider

Figure 1. Settings of the Interaction with the Virtual Patient.

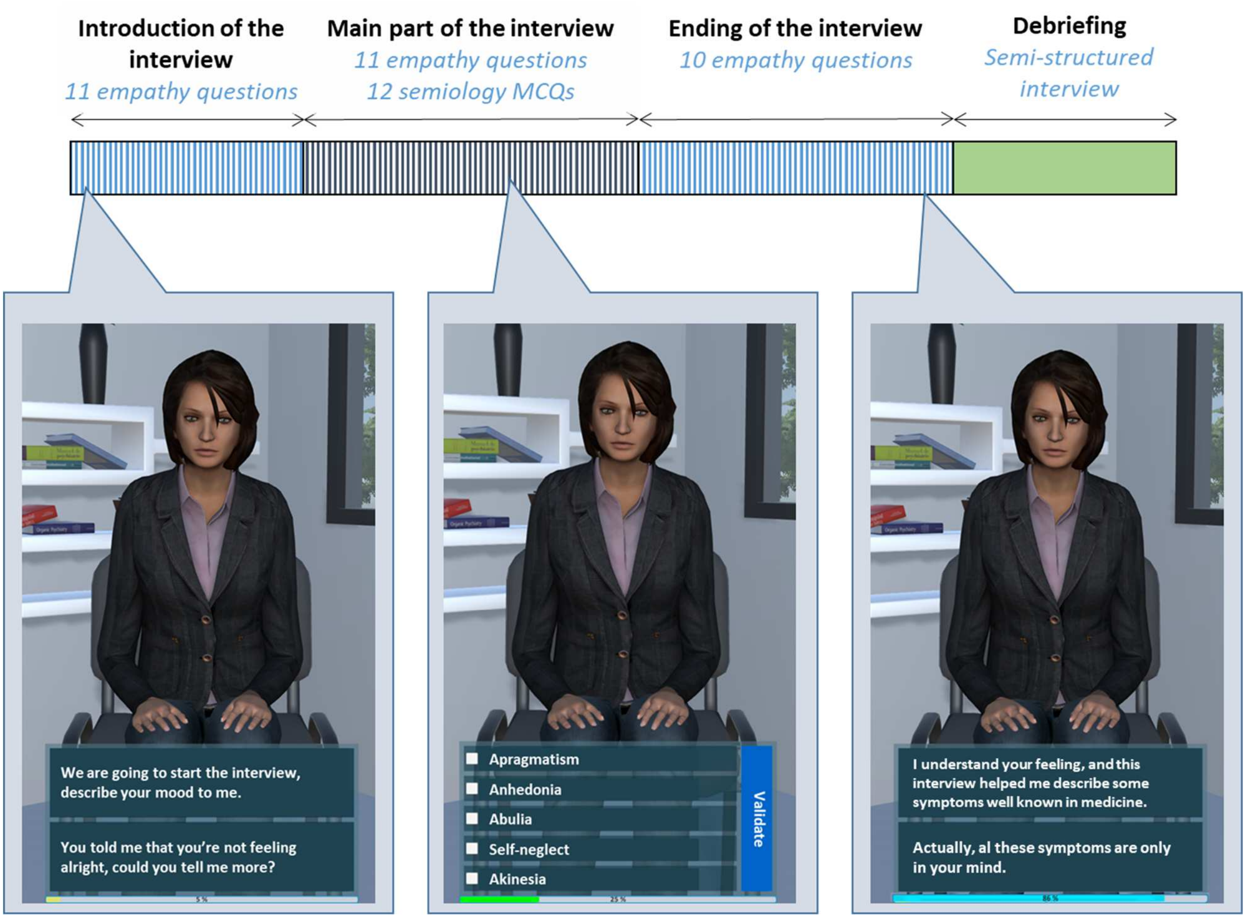
2.3 The VP interview

The interaction was based on a pre-determined scenario, with several options throughout the case and leading to a single endpoint (also known as a linear string of pearl narrative) (Huwendiek et al., 2009), in order to give the same information to all students. The scenario and questions were written by two experienced psychiatrists and followed the current psychiatry program at the medical school. The scenario led the student through the three phases of a psychiatric interview and lasted about the same duration as a real interview (around 35 min).

Repeatedly, the participant had to choose between two sentences the one that seemed the most appropriate to conduct the interview in an empathic manner (**Figure 2**). Particularly, the appropriateness of the question was based on simple and consensual rules in the field of psychiatric interview (Nordgaard et al., 2013; Shea, 2016):

- Avoid negative judgments (e.g., *"you are not trying hard enough"*)
- Prefer open questions (e.g. "*Now, could you describe your sleep?"* rather than "*Do you sleep well?")*
- Avoid multiple questions (e.g., *"Do you have allergies, a medical history, and do you take medication?")*
- Try to reformulate patients' answers in order to show them that their complaints are being taken into account (e.g., *"You told me that you feel like having a knot in your stomach, can you tell me more?"*)

When the sentence selected was not the appropriate one, the VP would answer saying that she did not understand or was a bit lost, and the accurate answer would be given to the student.



We are going to start the interview, describe your mood to me.

You told me that you're not feeling alright, could you tell me more?

□ Apragmatism
□ Anhedonia
□ Abulia
□ Self-neglect
□ Akinesia

I understand your feeling, and this interview helped me describe some symptoms well known in medicine.

Actually, al these symptoms are only in your mind.

Figure 2: Sequences of psychiatric interview and examples of empathy and semiology questions asked to user.

During the main part of the interview (i.e., when students had to gather symptoms listed in the DSM-5 to provide their diagnosis), semiology extraction was evaluated with multiple choice questions. Each question proposed five clinical signs and the participant had to select the one(s) demonstrated by the VP in the previous intervention (**Figure 2**). Once users had validated their answer, the system would give corrections, highlighting accurate and wrong answers to the user.

For each question, the participant had to read the answer(s) aloud in order to increase the realism of the interaction, and a vocal recognition module detected the selected answer.

2.4 Measures

2.4.1 Evaluation of psychiatric skills

2.4.1.1 Semiology extraction

Mirroring classical evaluation tools in medical exams, students had to select in 13 multi-choice questions (MCQs) the items corresponding to the depressive symptoms and pathological signs gathered during the interview. The number of right answers and errors was recorded, and participants received a score ranging from 0 to 20 (calculated from the raw score ranging from 0 to 65, corresponding to the total number of right answers).

2.4.1.2 Verbal empathy

Based on 32 two-choice questions, students had to choose the right answer promoting empathy and patient disclosure to elicit depressive symptoms. The number of right answers and errors was recorded, and participants received a score ranging from 0 to 20 (calculated from the raw score ranging from 0 to 32, corresponding to the total number of right answers).

2.4.1.3 Nonverbal empathy

In clinical practice, empathic skills are currently measured with paper-based scales, either self-reported by the physician (such as the Jefferson Scale of Physician Empathy; (Hojat et al., 2002)) or evaluated through observational scales (e.g., the Global Consultation Rating Scale; (Burt et al., 2014) or the Empathic Communication Coding System (Bylund and Makoul, 2005)). However, these measures are very dependent on the coder and do not capture variation in empathy over time. Therefore, we decided to assess objective and automatic emotion recognition by using emotion recognition software. During the interview, participants were video-recorded and their emotions were analyzed by Affectiva software[3].This system is based on face tracking to automatically detect emotions of individuals, following the Facial Action Coding System (FACS) (Friesen and Ekman, 1978). The emotion recognition software was first trained using human experts annotating hundreds of faces, then deep learning techniques were used to find common patterns between individuals to improve their solution as more persons are using it, and currently more than 7.5 million faces from 87 countries have been captured since the software was created (McDuff et al., 2016).

As a measure of the students' nonverbal empathic skills, we selected the Ekman's six basic emotions calculated by the software: *Joy*, *Fear*, *Anger*, *Disgust*, *Sadness*, and *Surprise*. These emotions are inferred based on distinct facial expressions, such as a smile for joy, or nose wrinkle and upper lip raised

[3] https://www.affectiva.com/

for disgust. For each emotion, the likelihood of an emotion is calculated, giving values ranging from 0 to 100 (details about how these dimensions are calculated can be found in their website[4]).

The software provides values every 3 milliseconds, enabling a time-based and objective evaluation of students' empathic skills during the interview. In addition, in order to identify empathic skills during the different phases of the interview, we annotated videos afterwards, indicating three different moments:

- *Questioning*: when the student was talking to the VP (i.e., analyzing and selecting the right empathic sentence to say to the VP);
- *Listening*: when the VP was talking (i.e., the student was listening and gathering symptoms);
- *Semiology MCQs*: when the participant was answering questions regarding the semiology of the VP.

We also annotated every moment that could interfere with face-tracking (e.g., the participant scratching his/her nose and therefore hiding a part of his/her face or turning around to ask something to the experimenter). All these moments were hence removed afterwards to clean the data and provide a more accurate measure.

2.4.2 User experience and attitude of students toward VP

After the interaction with the VP, a semi-structured debriefing was conducted by a psychiatrist, in order to go through the students' answers and errors committed, as well as to assess their attitudes toward the agent. Semi-structured interviews are well established techniques in a user-centered design as they enable hitherto unknown issues to be uncovered, provide flexibility and the possibility for clarification (Wilson, 2013). Open questions asked to the students were the following:

- What is your general opinion about the VP?
- What do you think about its usefulness for learning?
- What do you think about its credibility in terms of symptoms simulated?
- How acceptable was the duration of the interaction?
- How difficult were the questions?

The debriefing was video-recorded, and users' answers were transcribed and analyzed afterwards by the experimenters in terms of the lexical field of words repetitively used in the students' discourse.

2.5 Statistical analyses

[4] https://developer.affectiva.com/mapping-expressions-to-emotions/

Scores and errors were presented using means, standard deviations, minimum and maximum values. Group comparison analyses were performed with Student T-tests (for comparisons between two groups of subjects, here students' training department: Psychiatry vs. Neurology) and one-way ANOVA (for comparison between more than two groups, here between the three moments of the interaction: Questioning *vs.* Listening *vs.* Semiology MCQs). When ANOVAs suggested significant differences between groups, post-hoc analyses were performed to compare one group with another. Depending on the homogeneity of variances (screened with Levene's test), we used the Tukey post-hoc test (when variances were homogeneous) or the Games-Howell post-hoc test (when variances were heterogeneous). Pearson correlation analyses were performed to seek relations between emotion and engagement expressions, and scores and errors in communication and semiology questions. All statistical analyses were performed using SPSS software (version 18, PASW Statistics).

3. Results

3.1 Semiology extraction and verbal empathy: scores and errors

Globally, students had very good scores and made few errors (**Table 1**). Scores were significantly lower for semiology MCQs than for empathy questions (t(68) = 3.489; p < .001). Furthermore, students made significantly more errors during semiology MCQs than in empathy questions (t(68) = 8.064; p <.001).

Table 1: Descriptive statistics of scores and errors in empathy questions and semiology MCQs for all students

	Mean (SD)	[Min – max]
Total score	17.88 (0.77)	[15.87 – 19.37]
Verbal empathy - scores	18.41 (0.99)	[14.19 – 20]
Verbal empathy - errors	2.46 (1.54)	[0 – 9]
Semiology MCQs - scores	17.34 (1.01)	[14.77 – 19.38]
Semiology MCQs - errors	8.63 (3.23)	[2 – 17]

Note: SD: standard deviations

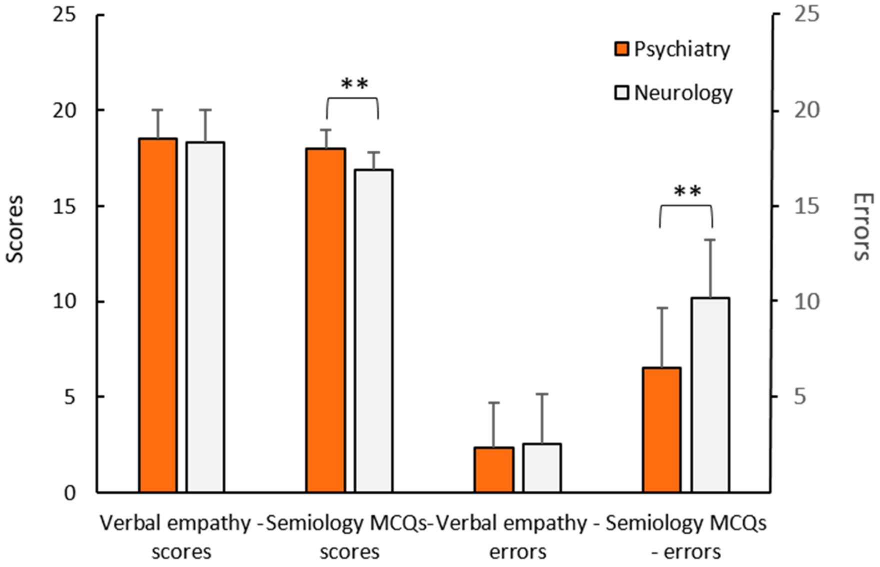


Figure 3. Scores and errors of verbal empathy questions and semiology MCQs for the two groups of participants. Error bars represent standard errors. ** p < 0.01

In addition, results showed significant differences regarding semiology MCQ scores and errors depending on the student's specialty department (i.e., neurology and psychiatry) (**Figure 3**), trainees in psychiatry having significantly better scores and making fewer errors than those trained in neurology (t(33) = 2.94; p = .006). No significant differences between specialty were found regarding empathy questions (p = .762).

3.2 Nonverbal empathy: emotion recognition data

Due to technical and face-tracking issues, emotion recognition was available for 21 subjects. Results showed very low values in every dimension measured, despite quite a high variation from one subject to another (**Figure 4**). ANOVA results showed a significant influence of the moment (i.e., Questioning, Listening, Semiology MCQs) on disgust (F(2,60) = 9,42; p < 0.001), surprise (F(2,60) = 11,05; p < 0.001) and fear (F(2,60) = 4,31; p = 0.018). Post-hoc tests suggested that students expressed more disgust during MCQs (p = 0.001) and questioning (p = 0.002) than during listening. The students also expressed more surprise during MCQs (p = 0.044) than during listening. Finally, data suggests that students express more fear during listening moments compared to MCQs (p = 0.049) and questioning moments (p = 0.027), but due to extremely low values (lower than 1 over a total score of 100), we believe that this result is not significant. In addition, correlation analyses showed that disgust values were significantly correlated with errors in MCQs (r = .46; p = .034) and scores in MCQs (r = -.46; p = .034), indicating that more errors during semiology MCQs (and therefore lower scores) were associated with more expression of disgust. Correlations between other measures (i.e., other emotions and other scores and errors) remained non-significant (p > 0.05).

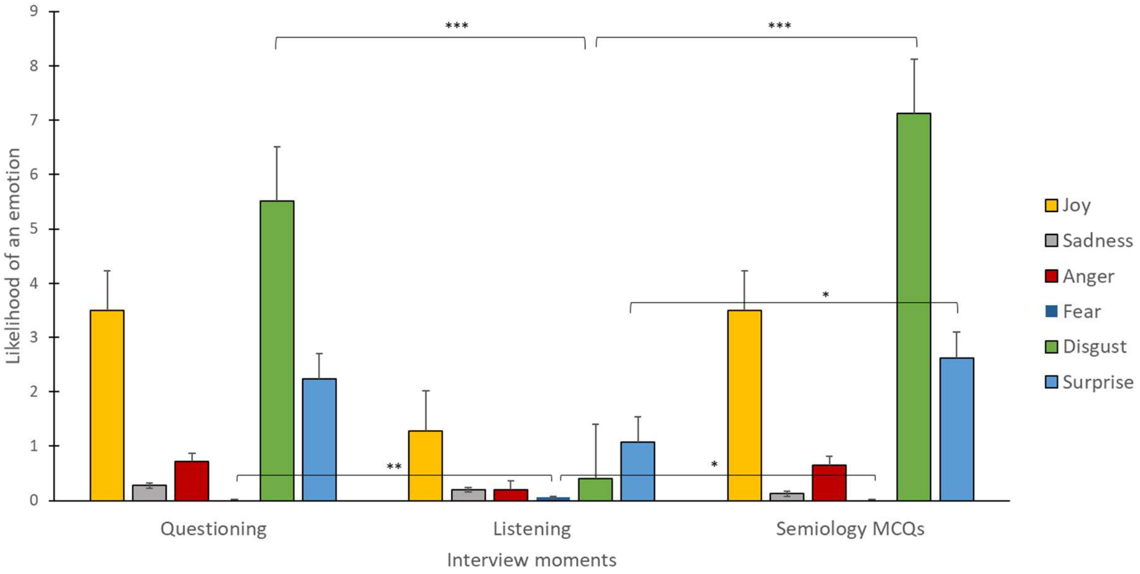


Figure 4. Average scores during questioning, listening and semiology MCQ moments

Error bars represent standard errors. * p < 0.05, ** p < 0.01, *** p < 0.001

3.3 Qualitative evaluation of VP by students

Generally, feedback given by the students after the interaction with the VP was very positive. Three main advantages were highlighted:

Pedagogic usefulness. Many students mentioned the benefits of the VP for learning, as it "*presents a good panel of symptoms*" (P1), "*uses the actual terms of the psychiatry manual*" (P22) and "*enables us to test [our] knowledge*" (P21). They also drew attention to the additional communicational skills learned during the interaction with the VP, such as "*learn how to conduct an interview*" (P6), "*[understand] which questions to ask a patient*" (P21). Moreover, they stressed the advantages of using digital solutions: *"they give you ready access to patients*" (P4), "*we cannot do internships in every domain*" (P24) and "*they could be used at home to prepare for an exam*" (P3).

User experience. Students expressed positive feelings regarding their interaction with the device, in terms of ease of use ("*not too difficult*" (P6, P7)), time consumption ("*not too long*" (P1, P3, P6, P17), "*half an hour, it's OK […], it is the same duration as a real interview*" (P7)) and enjoyment (e.g., "*awesome*" (P4), "*funny*" (P8, P17), "*cool*" (P14, P24), "*interesting*" (P7, P18, P20, P21), "*unexpected*" (P12), "*I wasn't expecting it to be that good!*" (P16), "*the patient is truly endearing!*" (P3).

Realism of interaction. Several participants mentioned the realism of the VP in terms of "*gestures*" (P4), "*sight*" (P12), and "*voice*" (P24). They found the interaction to be "*immersive*" (P25), and "*credible*" (P5): *"[depressive patients] are exactly like that!*" (P16), "*feels like conducting a real interview*" (P11). One student who did an internship in psychiatry even said *"I saw some real depressive patients, and they talk just like that. And psychiatrists ask the exact same questions!"* (P15).

They also pointed out some limitations. Notably, many found that the empathic questions (two-choice questions) were too easy: "*I felt like it was too obvious*" (P24), "*we understood quickly which question to choose*" (P20), "*two choices is too easy*" (P16), and "*too repetitive*" (P17) "*all the time the same type of questions*" (P15). However, as P26 said: "*the questioning was a bit obvious, but not when we moved to the semiology questions…*". Indeed, some complained about the difficulty of the questions listing psychiatric signs, as some terms (e.g., abulia, apragmatism, bradipsychia) might be complex and very specific: "*hard to remember it all*" (P25) "*I did not know all the semiologic terms*" (P20). Additionally, not all students had the same theoretical background regarding these terms: "*we have not learned about it yet*" (P17), "*we just started to see it in lectures*" (P14).

Finally, two students offered ideas for future work: "*It would be nice to have it for other disorders*" (P22), and "*it would be fun to do it with somebody undergoing a manic episode or something like that!*" (P9).

4. Discussion

This study is the first to validate the use of a virtual patient (VP) simulating a realistic psychiatric interview to train and assess medical students' semiology extraction and empathic skills in the field of affective disorders. The findings are encouraging and pave the way for new training modalities in psychiatric education.

The students managed to interact appropriately with the system, as overall they had good scores and made few errors. Interestingly, while both groups of students showed similar performance regarding the empathy questions, students having trained in psychiatry had better scores than their counterparts in neurology regarding the semiology extraction. To our knowledge, no study has used VPs to measure semiology extraction skills. The findings are therefore important, as they suggest that VPs can accurately measure psychiatric knowledge, and that the questions contain an appropriate level of difficulty to discriminate students' clinical knowledge.

Our VP was able to apply psychiatric interview recommendations effectively (Nordgaard et al., 2013; Shea, 2016) in order to provide realistic and practical training. Most studies involving VPs provide students only with text-based interfaces and short-time interaction situations (Deladisma et al., 2007; Foster et al., 2016; Maicher et al., 2019; Ochs et al., 2019), while this work shows that a more realistic interaction situation in terms of time and interface is feasible and applicable. Our VP also enabled to computerize the examination tools currently used, mainly based on paper-based multiple-choice examinations and human observations (Bhugra et al., 2017), making it more time-efficient and standardized for medical education.

Additionally, the objective measure of non-verbal empathy, based on emotion recognition software, showed appropriate discrimination between the different moments in the interview: questioning, listening and answering semiology MCQs. Results suggest that students remained neutral (*i.e.,* keep an emotional distance) when questioning and listening the VP, while they let their emotion be expressed when answering semiology questions. When answering MCQs, they could see their right and wrong answers, which might be the reason why they showed positive or negative emotions at that time. This was corroborated by the significant correlations between the number of errors they made and their expression of disgust, suggesting that the more students make errors, the more they express disgust. Another explanation could be that, based on how disgust values are calculated by the emotion recognition software (presence of noise wrinkle and raised upper lip), the expression of disgust might actually be closer to signs of concentration. On a higher level, our work gives insights about how to measure objectively both verbal and nonverbal expression of empathy during psychiatric interviews in real time, paving the way for new assessment tools in psychiatry and education research.

Lastly, during the debriefing sessions with the semi-structured interview, the students gave much positive feedback regarding the VP, as they understood its usefulness for pedagogy, shared their positive experience with the tool, which was seen as "*interesting*" and "*cool*", and underlined the perceived realism and credibility of the VP. In reference to well-known factors in the Human-Computer Interaction literature (*i.e.,* usefulness, ease of use, enjoyment), such feedback reflects good acceptance of the system and suggests that it will be readily used in the future (Hassenzahl, 2008; Venkatesh and Davis, 2000). In future studies, we could complement these debriefing sessions with standardized questionnaires to collect a more quantitative measure of acceptance of digital tools for mental health, as in other studies (Micoulaud-Franchi et al., 2016; Philip et al., 2019).

The debriefing sessions also revealed concerns of the students. First, the students found the questions dealing with empathy too easy, and sometimes even obvious and repetitive. Indeed, since we were testing a prototype, we wanted to apply recommendations for appropriate psychiatric interviews (Nordgaard et al., 2013; Shea, 2016) and decided to provide only two choices, which gave rise to some quite stereotypical questions. New versions could propose more than two choices, and subtler questions. However, it should be noted that only 3 students (8.57%) obtained the maximum score, suggesting that the answers to the empathy questions were not that obvious. On the other hand, some students reported concerns about the semiology MCQs being too difficult, suggesting possible adaptations, perhaps by proposing several levels of difficulty. This has been shown in the Human-Computer Interaction literature to increase user motivation (Oinas-Kukkonen and Harjumaa, 2009).

Additionally, the study suffers from some technological weaknesses. First, as highlighted above, the use of emotion recognition software to measure the expression of emotions could be questionable, as individuals' can be more complex (e.g., concentration instead of disgust). Also, face recording might have been disturbed by students' movements during the interaction (not in front of the camera, hand on their chin, scratching their nose, etc.), thereby inducing false positive errors. We tried to counteract this limitation by re-watch and annotate videos afterwards, but our data could still be questionable. An improvement could be to add gesture and voice recognition to gather more information about students' expression of emotions. Finally, the relatively simple interaction scenario, with only one end point and a few alternative scenes, led some students to underscore the redundancy and the obviousness of the empathy questions. Future versions could get closer to a real interaction with a patient, *e.g.* by increasing the complexity of the scenario or letting the students formulate their own questions (as designed by Kenny et al., 2008).

Together with system improvements, our work provides opportunities for future studies. First, analysis should focus on the assessment of VPs versus other assessment tools (e.g., the Jefferson Scale of Physician Empathy (Hojat et al., 2002) or the Empathic Communication Coding System (Bylund and

Makoul, 2005)) in order to ensure its accuracy. Second, in order to demonstrate the validity of VPs for training and evaluation, we should conduct longitudinal study that measure students' improvement when training with VPs, and their ability to transfer their skills from virtual reality to interaction with SPs (as in Foster et al., 2016) or with real patients, compared to classical medical training. Thirdly, an interesting application of VPs as a training tool could be to measure its influence on psychiatry stigmatization. Indeed, the field of psychiatry has a rather negative reputation with medical students, which makes psychiatry an unpopular career choice and impacts the treatment of mental illnesses (Lyons and Janca, 2015; Shen et al., 2014; Simon and Verdoux, 2018). Studies have shown that clerkship training (i.e., supervised clinical practice by 4th-year medical students) lowers negative attitudes towards psychiatry and increases students' decision to choose psychiatry as a career. A future study could therefore measure the impact of training with a VP on psychiatry stigmatization, compared to real practice or lecture training. Fourthly, VPs could be used to simulate other mental disorders, opening the way to a new field of research aiming at the following: i) proposing the precise identification and modeling of phenomenological features of symptoms of a disorder; ii) better simulating these features; iii) improving the realism of the symptomatology in the simulated interaction. Such methods could be very interesting to simulate subtle non-verbal manifestations of mental disorders such as the changes in prosody observed in depression (Cohn et al., 2009) or bizarreness in schizophrenia (Cermolacce et al., 2010; Gozé et al., 2019).

Taken together, our results pave the way for new digital tools to train and assess medical students conducting psychiatric interviews, making it possible to improve the difficult diagnosis of affective disorders by future physicians. By introducing and validating these new training tools, future psychiatrists should become trained in a new healthcare delivery model that is more patient-centered and integrated in the rapidly evolving field of psychiatry.

Author disclosure

Declaration of interest: None.

Contributors:

P.P. is the principal investigator in charge of the study

P.D., B.A., E.C., E.D.S., J.A.M., and P.P. designed the study and wrote the protocol

J.A.M., B.A. and P.P. wrote the psychiatric interview scenario

E.D.S developed and tested the virtual patient

P.D., B.A., and E.C. recruited the medical students

E.D.S., H.C., and J.A.M. ran the protocol and acquired the data for the study

L.D. and O.B. performed the statistical analyses on the collected data

L.D., E.D.S., O.B. and J.A.M. wrote the manuscript

All authors critically reviewed, edited the manuscript and approved the final version.

Funding: This project was supported by the French State within the framework of the national grants LABEX BRAIN (ANR-10-LABX-43), EQUIPEX PHENOVIRT (ANR-10-EQPX-01), and IdEx Bordeaux (ANR-10-IDEX-03-02).

Acknowledgments: We thank the medical students who took time to test our VP and give their precious feedbacks. We also thank Léa Chapolin, Caroline Louria and Anthony Latrille who helped with the video analysis, Emilien Bonhomme who participated in the development of the virtual patient, and Mr. Ray Cooke for proofreading.

References

American Psychiatric Association (Ed.), 2013. Diagnostic and Statistical Manual of Mental Disorders (DSM-5®). American Psychiatric Pub.

Batt-Rawden, S.A., Chisolm, M.S., Anton, B., Flickinger, T.E., 2013. Teaching Empathy to Medical Students: An Updated, Systematic Review. Acad. Med. 88, 1171. https://doi.org/10.1097/ACM.0b013e318299f3e3

Bhugra, D., Tasman, A., Pathare, S., Priebe, S., Smith, S., Torous, J., Arbuckle, M.R., Langford, A., Alarcón, R.D., Chiu, H.F.K., First, M.B., Kay, J., Sunkel, C., Thapar, A., Udomratn, P., Baingana, F.K., Kestel, D., Ng, R.M.K., Patel, A., Picker, L.D., McKenzie, K.J., Moussaoui, D., Muijen, M., Bartlett, P., Davison, S., Exworthy, T., Loza, N., Rose, D., Torales, J., Brown, M., Christensen, H., Firth, J., Keshavan, M., Li, A., Onnela, J.-P., Wykes, T., Elkholy, H., Kalra, G., Lovett, K.F., Travis, M.J., Ventriglio, A., 2017. The WPA- Lancet Psychiatry Commission on the Future of Psychiatry. Lancet Psychiatry 4, 775–818. https://doi.org/10.1016/S2215-0366(17)30333-4

Bromet, E., Andrade, L.H., Hwang, I., Sampson, N.A., Alonso, J., de Girolamo, G., de Graaf, R., Demyttenaere, K., Hu, C., Iwata, N., Karam, A.N., Kaur, J., Kostyuchenko, S., Lépine, J.-P., Levinson, D., Matschinger, H., Mora, M.E.M., Browne, M.O., Posada-Villa, J., Viana, M.C., Williams, D.R., Kessler, R.C., 2011. Cross-national epidemiology of DSM-IV major depressive episode. BMC Med. 9, 90. https://doi.org/10.1186/1741-7015-9-90

Burt, J., Abel, G., Elmore, N., Campbell, J., Roland, M., Benson, J., Silverman, J., 2014. Assessing communication quality of consultations in primary care: initial reliability of the Global

Consultation Rating Scale, based on the Calgary-Cambridge Guide to the Medical Interview. BMJ Open 4, e004339. https://doi.org/10.1136/bmjopen-2013-004339

Bylund, C.L., Makoul, G., 2005. Examining Empathy in Medical Encounters: An Observational Study Using the Empathic Communication Coding System. Health Commun. 18, 123–140. https://doi.org/10.1207/s15327027hc1802_2

Cassell, J., Sullivan, J., Churchill, E., Prevost, S., 2000. Embodied Conversational Agents. MIT Press.

Cermolacce, M., Sass, L., Parnas, J., 2010. What is Bizarre in Bizarre Delusions? A Critical Review. Schizophr. Bull. 36, 667–679. https://doi.org/10.1093/schbul/sbq001

Cohn, J.F., Kruez, T.S., Matthews, I., Yang, Y., Nguyen, M.H., Padilla, M.T., Zhou, F., De la Torre, F., 2009. Detecting depression from facial actions and vocal prosody, in: 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops. Presented at the 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, pp. 1–7. https://doi.org/10.1109/ACII.2009.5349358

Colliver, J.A., Conlee, M.J., Verhulst, S.J., Dorsey, J.K., 2010. Reports of the Decline of Empathy During Medical Education Are Greatly Exaggerated: A Reexamination of the Research. Acad. Med. 85, 588. https://doi.org/10.1097/ACM.0b013e3181d281dc

Cook, D.A., Erwin, P.J., Triola, M.M., 2010. Computerized Virtual Patients in Health Professions Education: A Systematic Review and Meta-Analysis. Acad. Med. 85, 1589. https://doi.org/10.1097/ACM.0b013e3181edfe13

Deladisma, A.M., Cohen, M., Stevens, A., Wagner, P., Lok, B., Bernard, T., Oxendine, C., Schumacher, L., Johnsen, K., Dickerson, R., Raij, A., Wells, R., Duerson, M., Harper, J.G., Lind, D.S., 2007. Do medical students respond empathetically to a virtual patient? Am. J. Surg. 193, 756–760. https://doi.org/10.1016/j.amjsurg.2007.01.021

Foster, A., Chaudhary, N., Kim, T., Waller, J.L., Wong, J., Borish, M., Cordar, A., Lok, B., Buckley, P.F., 2016. Using Virtual Patients to Teach Empathy: A Randomized Controlled Study to Enhance Medical Students' Empathic Communication. Simul. Healthc. 11, 181. https://doi.org/10.1097/SIH.0000000000000142

Friesen, E., Ekman, P., 1978. Facial action coding system: a technique for the measurement of facial movement. Palo Alto 3.

Gozé, T., Moskalewicz, M., Schwartz, M.A., Naudin, J., Micoulaud-Franchi, J.-A., Cermolacce, M., 2019. Reassessing "Praecox Feeling" in Diagnostic Decision Making in Schizophrenia: A Critical Review. Schizophr. Bull. 45, 966–970. https://doi.org/10.1093/schbul/sby172

Hassenzahl, M., 2008. The Interplay of Beauty, Goodness, and Usability in Interactive Products. Hum-Comput Interact 19, 319–349. https://doi.org/10.1207/s15327051hci1904_2

Hojat, M., Gonnella, J.S., Nasca, T.J., Mangione, S., Vergare, M., Magee, M., 2002. Physician Empathy: Definition, Components, Measurement, and Relationship to Gender and Specialty. Am. J. Psychiatry 159, 1563–1569. https://doi.org/10.1176/appi.ajp.159.9.1563

Huwendiek, S., leng, B.A.D., Zary, N., Fischer, M.R., Ruiz, J.G., Ellaway, R., 2009. Towards a typology of virtual patients. Med. Teach. 31, 743–748. https://doi.org/10.1080/01421590903124708

Kaplan, H.I., Sadock, B.J., 1988. Synopsis of psychiatry: Behavioral sciences clinical psychiatry, 5th ed, Synopsis of psychiatry: Behavioral sciences clinical psychiatry, 5th ed. Williams & Wilkins Co, Baltimore, MD, US.

Kenny, P., Parsons, T.D., Gratch, J., Rizzo, A.A., 2008. Evaluation of Justina: A Virtual Patient with PTSD, in: Prendinger, H., Lester, J., Ishizuka, M. (Eds.), Intelligent Virtual Agents, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 394–408.

Kim, S.S., Kaplowitz, S., Johnston, M.V., 2004. The Effects of Physician Empathy on Patient Satisfaction and Compliance. Eval. Health Prof. 27, 237–251. https://doi.org/10.1177/0163278704267037

Kleinsmith, A., Rivera-Gutierrez, D., Finney, G., Cendan, J., Lok, B., 2015. Understanding empathy training with virtual patients. Comput. Hum. Behav. 52, 151–158. https://doi.org/10.1016/j.chb.2015.05.033

Lyons, Z., Janca, A., 2015. Impact of a psychiatry clerkship on stigma, attitudes towards psychiatry, and psychiatry as a career choice. BMC Med. Educ. 15, 34. https://doi.org/10.1186/s12909-015-0307-4

Maicher, K.R., Zimmerman, L., Wilcox, B., Liston, B., Cronau, H., Macerollo, A., Jin, L., Jaffe, E., White, M., Fosler-Lussier, E., Schuler, W., Way, D.P., Danforth, D.R., 2019. Using virtual standardized patients to accurately assess information gathering skills in medical students. Med. Teach. 41, 1053–1059. https://doi.org/10.1080/0142159X.2019.1616683

McDuff, D., Mahmoud, A., Mavadati, M., Amr, M., Turcot, J., Kaliouby, R. el, 2016. AFFDEX SDK: A Cross-Platform Real-Time Multi-Face Expression Recognition Toolkit, in: Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems, CHI EA ’16. ACM, New York, NY, USA, pp. 3723–3726. https://doi.org/10.1145/2851581.2890247

Micoulaud-Franchi, J.-A., Quiles, C., Batail, J.-M., Lancon, C., Masson, M., Dumas, G., Cermolacce, M., 2018. Making psychiatric semiology great again: A semiologic, not nosologic challenge. L’Encéphale 44, 343–353. https://doi.org/10.1016/j.encep.2018.01.007

Micoulaud-Franchi, J.-A., Sauteraud, A., Olive, J., Sagaspe, P., Bioulac, S., Philip, P., 2016. Validation of the French version of the Acceptability E-scale (AES) for mental E-health systems. Psychiatry Res. 237, 196–200. https://doi.org/10.1016/j.psychres.2016.01.043

Nordgaard, J., Sass, L.A., Parnas, J., 2013. The psychiatric interview: validity, structure, and subjectivity. Eur. Arch. Psychiatry Clin. Neurosci. 263, 353–364. https://doi.org/10.1007/s00406-012-0366-z

Ochs, M., Mestre, D., de Montcheuil, G., Pergandi, J.-M., Saubesty, J., Lombardo, E., Francon, D., Blache, P., 2019. Training doctors’ social skills to break bad news: evaluation of the impact of virtual environment displays on the sense of presence. J. Multimodal User Interfaces 13, 41–51. https://doi.org/10.1007/s12193-018-0289-8

Oinas-Kukkonen, H., Harjumaa, M., 2009. Persuasive Systems Design: Key Issues, Process Model, and System Features. Commun. Assoc. Inf. Syst. 24. https://doi.org/10.17705/1CAIS.02428

Philip, P., Dupuy, L., Auriacombe, M., Serre, F., de Sevin, E., Sauteraud, A., Micoulaud-Franchi, J.-A., 2019. Trust and acceptance of a virtual psychiatric interview between embodied conversational agents and outpatients. Npj Digit. Med.

Plakun, E.M., 2015. Psychotherapy and Psychosocial Treatment: Recent Advances and Future Directions. Psychiatr. Clin. 38, 405–418. https://doi.org/10.1016/j.psc.2015.05.012

Shea, S.C., 2016. Psychiatric Interviewing E-Book: The Art of Understanding: A Practical Guide for Psychiatrists, Psychologists, Counselors, Social Workers, Nurses, and Other Mental Health Professionals. Elsevier Health Sciences.

Shen, Y., Dong, H., Fan, X., Zhang, Z., Li, L., Lv, H., Xue, Z., Guo, X., 2014. What Can the Medical Education Do for Eliminating Stigma and Discrimination Associated with Mental Illness among Future Doctors? Effect of Clerkship Training on Chinese Students’ Attitudes. Int. J. Psychiatry Med. 47, 241–254. https://doi.org/10.2190/PM.47.3.e

Silverman, J.J., Galanter, M., Jackson-Triche, M., Jacobs, D.G., Lomax, J.W., Riba, M.B., Tong, L.D., Watkins, K.E., Fochtmann, L.J., Rhoads, R.S., Yager, J., 2015. Practice Guidelines for the

Psychiatric Evaluation of Adults. Am. J. Psychiatry 172, 798–802. https://doi.org/10.1176/appi.ajp.2015.1720501

Simon, N., Verdoux, H., 2018. Impact de la formation théorique et clinique sur les attitudes de stigmatisation des étudiants en médecine envers la psychiatrie et la pathologie psychiatrique. L'Encéphale 44, 329–336. https://doi.org/10.1016/j.encep.2017.05.003

Tolks, D., Schäfer, C., Raupach, T., Kruse, L., Sarikas, A., Gerhardt-Szép, S., Kllauer, G., Lemos, M., Fischer, M.R., Eichner, B., Sostmann, K., Hege, I., 2016. An Introduction to the Inverted/Flipped Classroom Model in Education and Advanced Training in Medicine and in the Healthcare Professions. GMS J. Med. Educ. 33. https://doi.org/10.3205/zma001045

Venkatesh, V., Davis, F.D., 2000. A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. Manag. Sci. 46, 186–204. https://doi.org/10.1287/mnsc.46.2.186.11926

Wilson, C., 2013. Interview Techniques for UX Practitioners: A User-Centered Design Method. Newnes.