Can Psychotherapy Trainees Distinguish Standardized Patients From Real Patients?

A Pilot Study

Destina Sevde Ay-Bryson¹, Florian Weck¹, Peter Eric Heinze¹, Thomas Lang,^{2,3} and Franziska Kühne¹

¹Clinical Psychology and Psychotherapy, University of Potsdam, Germany

²Psychology and Methods, Jacobs University Bremen, Germany

³Christoph-Dornier-Stiftung, Bremen, Germany

Abstract: *Background:* Under the new psychotherapy law in Germany, standardized patients (SPs) are to become a standard component in psychotherapy training, even though little is known about their authenticity. *Objective:* The present pilot study explored whether, following an exhaustive two-day SP training, psychotherapy trainees can distinguish SPs from real patients. *Methods:* Twenty-eight psychotherapy trainees ($M = 28.54$ years of age, $SD = 3.19$) participated as blind raters. They evaluated six video-recorded therapy segments of trained SPs and real patients using the Authenticity of Patient Demonstrations Scale. *Results:* The authenticity scores of real patients and SPs did not differ ($p = .43$). The descriptive results indicated that the highest score of authenticity was given to an SP. Further, the real patients did not differ significantly from the SPs concerning perceived impairment ($p = .33$) and the likelihood of being a real patient ($p = .52$). *Conclusions:* The current results suggest that psychotherapy trainees were unable to distinguish the SPs from real patients. We therefore strongly recommend incorporating training SPs before application. Limitations and future research directions are discussed.

Keywords: evidence-based training, learning, simulated patients, simulation-based education, therapist competence

Können Psychotherapeut_innen in Ausbildung standardisierte Patient_innen von realen Patient_innen unterscheiden? Ergebnisse einer Pilotstudie

Zusammenfassung: *Theoretischer Hintergrund:* Mit dem neu eingeführten Direktstudium für zukünftige Psychotherapeut_innen (PiA) wird der Einsatz von standardisierten Patient_innen (SP) in der Lehre zunehmen, obwohl die Authentizität der Rollendarstellungen durch SP empirisch bislang kaum untersucht wurde. Ziel der vorliegenden Studie war es daher zu untersuchen, ob SP trainiert werden können, dass Psychotherapeut_innen in Ausbildung (PiA) SP von realen Patient_innen nicht unterscheiden können. *Methode:* Insgesamt nahmen 28 PiA ($M = 28.54$ Jahre, $SD = 3.19$) als verblindete Rater teil. Sie haben sechs Therapiesitzungen von trainierten SP und realen Patient_innen mit der Skala Authentizität von Patientendarstellungen bewertet. *Ergebnisse:* Die Authentizitätswerte von SP unterschieden sich nicht signifikant von realen Patient_innen ($p = .43$). Deskriptive Ergebnisse legen nahe, dass ein SP im Schnitt am authentischsten bewertet wurde. Darüber hinaus unterschieden sich SP und reale Patient_innen nicht hinsichtlich der wahrgenommenen Beeinträchtigung ($p = .33$) sowie der Wahrscheinlichkeit, als reale/r Patient_in bewertet zu werden ($p = .52$). *Fazit:* Die vorliegenden Ergebnisse legen nahe, dass PiA SP von realen Patient_innen nicht unterscheiden konnten. Daher legen wir ein ausführliches Training der SP nahe, bevor sie für Studium und Lehre eingesetzt werden. Die Limitationen sowie zukünftige Forschungsideen werden diskutiert.

Schlüsselwörter: evidenzbasiertes Training, Lernen, Simulationspatient_innen, simulationsbasierte Lehre, therapeutische Kompetenz

According to the new psychotherapy law in Germany (Psychotherapeutenausbildung, 2019), standardized patients (SPs) are to become a standard component in psychotherapy training similar to that found in medical training, where SPs have been present for decades as part of the objective and structured clinical examination

(e.g., Adamo, 2003). From a theoretical point of view, a discussion is ongoing concerning the role of the assessment and training of therapist competence in the field of clinical psychology. In a recent review, Muse and McManus (2013) extended Miller's (1990) hierarchical framework for clinical assessment by measuring the

competence of cognitive behavioral therapists. According to this framework, different levels of clinical competence can be assessed in different ways. *Knowledge*, representing the first level on the hierarchy, may be assessed using essays or multiple-choice questions (Muse & McManus, 2013). Assessing *practical understanding* represents the second level, which may additionally be evaluated by case reports or short-answer clinical vignettes. The penultimate level on the hierarchy aims to answer the question of whether the therapist *demonstrates proper skills*, followed by the final level, *clinical practice*, which refers to the therapist's use of those skills. The penultimate level becomes especially relevant in psychotherapy training. To assess the practical application of knowledge – i.e., skills – Muse and McManus (2013) referred to the utilization of standardized role-playing, which they define as “artificial simulations of clinical scenarios in which a therapist interacts with an individual playing the role of a standardized patient” (p. 490).

SPs were first introduced by Barrows and Abrahamson (1964) to evaluate the clinical performance of students in clinical neurology, and today they are widely accepted in the medical field (Kühne et al., 2018). SPs are healthy laypersons who portray a clinical problem in a standardized manner. Although they have been regarded with skepticism in the context of psychotherapy training and research (Hodges et al., 2014; Kühne et al., 2020), they are currently being increasingly introduced into clinical psychology and psychotherapy (Kühne et al., 2020). For instance, a previous study (Partschfeld et al., 2013) investigated the integration of SPs in psychotherapy training and found promising effects regarding therapeutic skills. Further, under the new psychotherapy law, SPs will likely be integrated even more throughout Germany. While the use of SPs for the assessment of competence has shown promising potential for training and research, it has also been noted that it potentially simplifies clinical complexity and may hinder authenticity (Sharpless & Barber, 2009).

Because of increasing attention being paid toward SPs in research and training (Melluish et al., 2007; Partschfeld et al., 2013; Sheen et al., 2020), it is crucial to examine the authenticity of SPs if we are to overcome reservations associated with lack of realness in simulations. Here, we refer to the definition of authenticity proposed by Wünderlich et al. (2012), namely, the “impossibility of distinguishing SPs from (real) patients” (p. 501). Few studies have examined the authenticity of SPs by comparing them with real patients. In the often-cited pilot study by Krahn et al. (2002), SPs were provided with a case in an outline format, which enabled them to improvise their answers during simulated interactions. SPs received one training session before the actual simula-

tion, but the duration of the training was not specified. Students who conducted interviews with patients could correctly identify the SPs most of the time. The results of this study also showed that 91% of students who believed the patient to be an SP felt less empathy toward them. The authors of this pilot study concluded that “training must focus on facilitating actors' ability to convey emotion realistically and therefore evoke empathy in the interviewer” (p. 30).

In contrast, in a more recent study (Wünderlich et al., 2012), SPs were trained for about 4 hours in the authentic simulation of a patient. Then experienced psychiatrists were asked to retrospectively assign one of three labels (SP, patient, unsure) to each interviewed person. Their results demonstrated that, although SPs were rated as less authentic than real patients, in 70 of 114 (61.40%) SP cases they were not detected. Wünderlich and colleagues concluded that, “with proper training, [SPs] can reach a high level of authenticity in presenting major psychiatric disorders when rated by experienced psychiatrists” (p. 501). Although it seems plausible that training SPs would result in more authentic portrayals of patient cases (Partschfeld, 2013), the literature provides little evidence for this assumption. Consequently, there is no gold standard yet on training SPs.

Although both studies (Krahn et al., 2002; Wünderlich et al., 2012) examined the authenticity of SPs as evaluated by external judges, they both concentrate solely on subjective, single-item assessments (e.g., retrospective allocation to a group) and medically related samples (e.g., psychiatrists). Further, the simulated interactions were not standardized, which limits comparability.

Objective

In the current pilot study, we addressed these limitations and developed a two-day SP training. In a previous randomized-controlled study (Ay-Bryson et al., under review), we found that SPs can be trained to be more authentic using a detailed role-script on a patient case compared to basic information. In the current pilot study, we explored whether, following thorough training, SPs can be distinguished from real patients by psychotherapy trainees. We addressed the following research questions: (1) Can psychotherapy trainees distinguish between trained SPs and real patients? Further, since an SP can be evaluated as authentic (but still not be perceived as a real patient), we were interested in the following research questions: (2) Is there a relationship between authenticity and the likelihood of the interviewed person being a real patient? Is there a relationship between authenticity and their psychological impairment?

Methods

Study Design

Participants who were eligible to participate (i.e., raters; see below) watched six video-recorded simulations of 5-minute therapy segments. The raters were instructed to evaluate the persons interviewed in the simulations concerning their authenticity and psychological impairment. Participants were not informed about the status of the interviewed persons (i.e., whether real or SPs), nor were they informed that the videos would even include SPs and/or real patients, which was ensured by using the term “person” in the instruction.

Participants

Raters

A total of 29 raters participated in the current study, one of whom was excluded from analyses because of too much missing data (i.e., answered only half of the items). Hence, the final sample consisted of $N = 28$ persons (82.14% female). Anyone currently undergoing psychotherapy training or licensed psychotherapists was eligible to participate; all participants provided informed consent before participation. The mean age of the raters was 28.54 years ($SD = 3.19$; range = 24–40 years). Most of them were psychotherapy trainees ($n = 26$; 92.86%), all of whom were specialized in cognitive behavioral therapy. The average duration of their psychotherapy training was 8.73 months ($SD = 2.40$; range = 4–10 months).

Standardized and Real Patients

The mean age of the four female SPs was 22.75 years ($SD = 3.63$; range = 20–29) and they reported no prior experience as SPs before taking part in the current study. The SPs were part of an ongoing research project (Kühne et al., 2020) and were selected based on the following criteria: no mental illness, ability to reflect, and joy doing theatrical work. The two real patients aged 23 (female) and 24 (male), both diagnosed with depression (recurrent and episode, respectively), were undergoing treatment at the Outpatient Unit of the Department of Clinical Psychology and Psychotherapy at the University of Potsdam. We recruited these patients using the following questions: Do they fulfill the criteria for mild depression? Are they available? Do they consent to being video-recorded? Does their sociodemographic data match those of the SPs? The real patients received background information on the study as well as a description of the situation, namely, that the simulation would take place in the laboratory to

ensure the comparability of the simulated interactions. The therapy situation consisted of a therapist exploring a specific situation typical for depression reported by the SP / real patient. Accordingly, SPs portrayed a patient experiencing a first depressive episode. All SPs and real patients provided informed consent before participation.

Procedure

Our university's Ethics Committee approved the study (no. 9/2018, University of Potsdam).

Setting

The six video segments were recorded in the laboratory designed as a therapy room at the University of Potsdam (see Figure 1). To ensure comparability, all sessions were conducted by a licensed psychotherapist (FK). Four of the sessions featured a trained SP and two featured real patients, whereby the sessions focused on the exploration of a specific situation. The ratings by psychotherapy trainees took place at the Institute for Psychological Psychotherapy Training in Bremen (IPP) and at the University of Potsdam by psychotherapy trainees of the Psychological-Psychotherapeutic Institute (PPI) at the UP Transfer. All video segments were viewed sequentially. The order of the video segments and the duration of each video was the same for all raters.

SP Training

The training of the SPs to authentically simulate patient roles (i.e., indistinguishable from real patients) consisted of a two-day workshop at the University of Potsdam, which altogether took 12 hours, including homework. The training consisted of seven units: (1) project introduction, (2) SP concept, (3) psychoeducation, (4) role descriptions, (5) authenticity, (6) role analysis and comparability, and (7) role-play exercises, de-roling, and feedback. We conceptualized the SP training following previously published manuals (Brem et al., 2018; Scherer & Ehrhardt, 2017) as well as published overviews (Adamo, 2003; Peters & Thrien, 2018; Voderholzer, 2007).

During the *project introduction phase*, the accompanying research project (Kühne et al., 2020) was presented and all staff members were introduced. During the *SP concept phase*, the definition of an SP as well as their benefits and drawbacks for psychotherapy training were discussed. Subsequently, participants were made familiar with their tasks, including the scripted scenarios of a depressed patient. The *psychoeducation phase* included general information about mental disorders followed by a brief introduction to cognitive behavioral therapy (CBT). Then, the diagnostic criteria of depression were presented, and

Figure 1. Study design.

to further clarify the manifestation of depressive symptoms, a case example was discussed. Furthermore, participants were informed about treatment options for depression with a focus on CBT followed by information regarding particular therapeutic strategies, such as changing behavior and changing thoughts (Hautzinger, 2013). Six different *role descriptions* had to be read carefully at home by all participants, which were then discussed regarding the diagnostic criteria of the portrayed disorder and difficulties in the portrayal. Different aspects of *authenticity* were discussed (“In your opinion, what makes a portrayal authentic?”), which aimed to clarify the individual concepts of authenticity. It was further explained that authentic SPs present cases more typically than real patients, and that those representations are often more tangible for students or trainees. The discussion served to clarify the advantages of authentic SP role-plays as facilitators of learning conditions. Furthermore, to create an awareness of the symptoms typical for depression, specific diagnostic criteria were conveyed. To improve authenticity, video analysis of a role model and of the SPs themselves was implemented, and positive and negative performances were discussed (“How convincing did you find the SP?”). During *role analysis and comparability*, different role descriptions were analyzed by discussing questions about demographic facts, symptoms, social environment, biography, thoughts, and feelings of the scripted person. Participants were encouraged to think of strategies for carrying out an authentic role-

play of each character. Subsequently, it was discussed that the SPs should make an effort to simulate the patient in a consistent and hence comparable manner during the subsequent interactions. Finally, *role-play exercises, de-roling, and feedback* involved practicing basic acting skills. Acting exercises were conducted with a focus on non-verbal acting. This was followed by the introduction of de-roling techniques, such as sitting on a different chair, taking off clothes and requisites associated with the role, and discussing the latest session with a peer. Finally, role-plays were carried out and evaluated, and feedback was given by a licensed psychotherapist (FK) to the SPs.

Simulated Situations and Scenarios

The simulated situations were based on scripted scenarios. For the development of the scripts, we followed those previously published (Voderholzer, 2007). The scripts included information regarding depressive symptoms, factors that contribute to the development and maintenance of depression as well biographical aspects (for a scenario example, see Supplement 1; and Kühne et al., 2020). For instance, all SPs were to simulate a depressive patient with comparable symptoms and impairment. However, the background story of the SPs differed slightly in order not to make SPs too obvious compared to real patients. For example, one SP portrayed a person feeling overwhelmed by the question of what to do after school graduation, whereby another SP portrayed a person experiencing perinatal depression. The role-scripts

Table 1. Descriptive results on APD and additional items' ratings

Interviewed person number (status)	Authenticity	Real patient	Impairment
		<i>M (SD)</i>	
1 (SP)	1.96 (.43)	.36 (.22)	.40 (.21)
2 (real patient)	1.89 (.63)	.45 (.32)	.52 (.22)
3 (SP)	1.99 (.67)	.54 (.25)	.53 (.20)
4 (SP)	1.96 (.67)	.47 (.33)	.43 (.23)
5 (real patient)	2.13 (.59)	.65 (.28)	.40 (.18)
6 (SP)	2.48 (.45)	.69 (.24)	.62 (.21)
Group			
SPs	2.10 (.35)	.51 (.14)	.50 (.13)
Real patients	2.01 (.49)	.55 (.24)	.46 (.16)

Note. Authenticity was measured with the APD scale on a 4-point Likert scale: 0 = *strongly disagree*, 1 = *disagree*, 2 = *agree*, 3 = *strongly agree*. Items "real patient" and "impairment" were scored from 0% to 100% and reported in the table as normalized proportions. *M* = mean score; *SD* = standard deviation; SP = standardized patient.

were previously evaluated by an independent licensed psychotherapist and an actual depressive patient (not included in the study) regarding comprehensibility, feasibility, relevance, and suitability for an authentic portrayal of a depressive disorder.

Measures

Authenticity of Patient Demonstrations Scale (APD)

To assess the authenticity of SPs and real patients, we used the 10-item Authenticity of Patient Demonstrations scale (APD; Ay-Bryson et al., in press). One example item is: "The person describes disorder-specific symptoms appropriately," which was scored on a 4-point Likert scale ranging from 0 = *strongly disagree* to 3 = *strongly agree*. The APD had been demonstrated in a previous study (Ay-Bryson et al., in press) to distinguish well between authentic and unauthentic SPs, to have a one-factorial structure, and to have a good internal consistency (Cronbach's $\alpha = .83$). Further, there is evidence of its convergent and discriminant validity (Ay-Bryson et al., in press), as it correlated ($r = .82$) with an established tool for the assessment of SP performance in medical contexts and demonstrated no correlation with unrelated symptoms (i.e., anxiety) that were not supposed to be portrayed ($r = .004$). In the current study, Cronbach's α ranged from .84 to .95. The internal consistency is comparable to that of the original study of the scale (Ay-Bryson et al., in press) and can be considered good to excellent. Interrater reliability between all raters over the whole sample was $ICC_{(2,28)} = .76$, $p < .001$, 95% CI [.49, .94] (Koo & Li, 2016). According to Koo and Li (2016), ICCs between .75 and .90 indicate good reliability.

Additional Items

Further, we asked the raters to evaluate *how psychologically impaired* (i.e., impairment) each patient appeared. Finally, raters estimated *how likely* it was that each interviewed person was a *real patient* (i.e., real patient). Both items used an 11-point scale ranging from 0% to 100%.

Statistical Analyses

A final sample size of $N = 6$ simulated interactions based on $N = 28$ raters was included in statistical analyses. Two reversed-scored items of the APD were inversed. We then computed means over all raters for each SP / real patient and calculated group means for the comparison of SPs vs. real patients.

To test whether SPs and real patients differed, we computed two sample *t*-tests. To explore the relationship between authenticity and the additional two items, we computed Pearson's correlations and used the Bonferroni-corrected *p*-values (see Table 2). All statistical analyses were performed with R version 3.4.2 (2017); the alpha level was set to .05.

Results

Indistinguishability of SPs (Research Question 1)

Table 1 provides the mean scores of the APD and additional items as well as grouped means. On average, the raters "agreed" that the SP group ($M = 2.10$, $SD = .35$)

Table 2. Pearson's correlations based on ratings ($N = 28$)

Measure	1	2	3	4	5	6
	(95 % confidence interval)					
1 Authenticity RP	1					
2 Authenticity SP	.61** (.30, .80)	1				
3 Real patient RP	.76** (.53, .88)	.36 (-.02, .64)	1			
4 Real patient SP	.19 (-.19, .53)	.62** (.31, .80)	.27 (-.12, .58)	1		
5 Impairment RP	.13 (-.26, .48)	-.08 (-.44, .30)	.43 (.06, .69)	.10 (-.28, .46)	1	
6 Impairment SP	-.19 (-.53, .20)	.19 (-.20, .52)	-.19 (-.52, .20)	.52 (.18, .75)	.23 (-.16, .56)	1

Note. RP = real patient; SP = standardized patient. Bonferroni corrected p -values: * $p < .003$, ** $p < .001$.

as well as the real-patients group ($M = 2.01$, $SD = .49$) portrayed their roles authentically. The raters indicated a 51 % ($SD = .14$) probability for the SPs to be a real patient and a 55 % probability for the real patients. Finally, SPs were evaluated to have a 50 % (.13) psychological impairment, whereby the real patients were evaluated as slightly less impaired ($M = .46$ %, $SD = .16$).

Authenticity

The APD means of SPs did not differ significantly from the APD means of the real patients: $t(48.93) = -.79$, $p = .43$; Cohen's $d = -.21$, 95 % CI [-.75, .33].

Realness and Impairment

SPs and real patients did not differ significantly on the item "real patient": $t(43.79) = .66$, $p = .52$; Cohen's $d = .18$, 95 % CI [-.36, .71]. SPs and real patients did not differ significantly regarding the item "impairment": $t(51.92) = -.98$, $p = .33$; Cohen's $d = -.26$, 95 % CI [-.80, .28].

Correlations Between Authenticity, Realness, and Impairment (Research Question 2)

Table 2 presents the Pearson's correlations between the APD and the additional items. As expected, authenticity strongly correlated with the item "real" ($r = .62$ – $.76$, $p < .001$). On the other hand, the item "impairment" did not correlate significantly with authenticity ($r_{\text{realpatient}} = .13$, $p = .51$; $r_{\text{SP}} = .19$, $p = .35$). Further, there were strong but nonsignificant correlations between both additional items ($r_{\text{SP}} = .43$, $p = .02$; $r_{\text{realpatient}} = .52$, $p = .004$).

Discussion

To facilitate an increasing spread of SPs worldwide – but also particularly in Germany – the current paper suggests careful training of SPs before their deployment. The current pilot study explores whether psychotherapy trainees can distinguish between real patients and SPs who have been trained in an elaborate 2-day training. Furthermore, this study revealed no significant difference between real patients and SPs regarding authenticity, perceived impairment, and the perceived likelihood of being a real patient. Still, we caution against over-relying on the results and propose to use equivalence tests based on bigger sample sizes to determine the absence of an effect (Anderson & Maxwell, 2016; Lakens, 2017).

Pilot results from the current study suggest that psychotherapy trainees could not distinguish trained SPs from real patients. In fact, considering the descriptive results, one of the two real patients was rated the least authentic of all interviewed persons, and the person with the highest authenticity score happened to be an SP. Because of these results, we conclude that the training we conducted was effective in equipping SPs with the ability to portray a depressive patient authentically, perhaps because of the theoretical input regarding both the background and conceptualization of SPs and the practical phase of the training, during which SPs received feedback from a licensed psychotherapist (FK). SPs may also be able to better study their roles because they received an introduction to the theoretical background of the clinical picture of depression, which included general knowledge as well as specific diagnostic criteria and treatment options (Hautzinger, 2013). However, to draw firm conclusions on the SP training proposed in our study, we

suggest future studies to conduct pre-post training comparisons of authenticity.

In line with our expectation, persons who were rated as authentic were also evaluated as likely to be a real patient. We interpret this result as further evidence that the raters could not tell SPs from real patients. Although the correlation between psychological impairment and the likelihood of being a real patient was not significant after a Bonferroni correction, this relationship could be investigated in future studies based on bigger sample sizes. Arguably, perceived impairment plays a central role in the context of authenticity. Interestingly, in their study Wünderlich and colleagues (2012) found that SPs were rated significantly better for the items “case is typical” and “symptoms are obvious to students.” Since clinically significant impairment is one main criterion, it is conceivable that the introduction to specific diagnostic criteria is of major importance when training SPs.

Like previous studies (Krahn et al., 2002; Wünderlich et al., 2012), we compared trained SPs with real patients, though we adapted this approach to psychotherapy training. Notably, we ensured comparability between simulated interactions by standardizing the therapist who conducted all simulations. Finally, we considered psychotherapy trainees as raters, unlike the study of Wünderlich et al. (2012), in which experienced psychiatrists rated SPs. Arguably, the perspective of trainees is relevant considering that mostly they are to be trained and assessed in interactions with an SP. On the other hand, given that the training with SPs aims to prepare trainees for real clinical encounters, experienced clinicians should be consulted too when evaluating the representativity and authenticity of SPs. Ideally, future studies should incorporate both perspectives (i.e., trainees and experienced clinicians).

Considering average descriptive results of the item real, we noticed that trainees were rather unsure whether the interviewed person was a real patient. This may potentially be associated with the early stage of training as most raters were in their first year of training. On the other hand, those who evaluated the interviewed person as authentic also evaluated the person as likely to be a real patient, as demonstrated in the strong correlations. Therefore, during SP training it seems crucial to focus on the aspect of authenticity. An alternative explanation for the uncertainty may also be that, although we successfully trained SPs to be evaluated as authentic, there is still room for improvement in their performances. Higher levels of authenticity may, for instance, be reached through refresher training sessions, through inviting real patients to a training session, or by using model learning situations adjusted to the scenarios to be simulated. However, an SP may also be authentic but not believed to be a real patient. It would be interesting to investigate whether this would

contribute differently to learning effects. Future studies could examine whether the perceived authenticity has a stronger effect on learning than the perceived believability of the SP. Conceivably, trainees would still benefit from training with authentic SPs even if they are not believed to be real. If this were empirically the case, one consequence would be that masking the status of the SP would not be necessary for training purposes.

Limitations and Future Research Directions

While the results of the current pilot study are promising, limitations should be considered. Importantly, future studies should conduct a priori power analyses to determine the necessary sample size. Since the current pilot study comprised only a small sample size, the generalizability of the reported results is limited. Consequently, future studies should ideally consider a bigger sample size of both SPs and real patients. Another limitation of the present study is that the groups were unbalanced (i.e., 4:2 SPs:real patients) because of the patient availability at our outpatient clinic. One strength of the current study lay in our successfully matching the overall sociodemographic data of the real patients with those of our SPs, and that the clinical picture was comparable, too. Despite the clustered structure of our data (i.e., each psychotherapy trainee rated all patients), because of our sample size, we could not conduct analyses that consider multilevel data (Maas & Hox, 2005). Consequently, future studies would benefit from considering bigger sample sizes in this manner as well.

Further, we considered only one clinical picture, i.e., depression. Thus, the present results should be replicated using different mental health problems. Arguably, different mental illnesses may vary in the difficulty to portray them authentically. For instance, more complex mental disorders, such as mania or psychoses, would seem to be more difficult to portray (Kühne et al., 2018). Similarly, it may be more challenging to portray patients with comorbid disorders. Studying this, however, would be highly desirable as comorbidity is usually prevalent in clinical encounters (e.g., Andrews et al., 2001; Angst et al., 2004). It is conceivable that experienced SPs with longer histories of being engaged as an SP could more likely portray complex disorders more authentically. Moreover, both additional items, i.e., real and impairment, are single items and should therefore be interpreted with caution. Future studies should consider briefing raters beforehand to prevent potential confusion regarding these variables.

For economic reasons, we included 5-minute therapy segments on focused interactions. Thus, the results deliver

first clues but should be expanded in future studies. In particular, it would be relevant to replicate the present results based on longer segments or even full sessions.

Future studies may also benefit from conducting dismantling studies to identify which of the training components accounted for authenticity (e.g., Ay-Bryson et al., under review). Such results may be favorable for SP programs with less time and fewer resources to adapt the length of the training to availability.

Finally, raters of the current study were blind to the status of the interviewed persons – indeed, they were not informed that some of the patients would be SPs and some real patients, unlike in the study by Wünderich et al. (2012). The instruction given to the raters may have affected how they perceive the interviewed persons. When conducting observational studies, we should consider potential biases, such as the halo or contrast effect, that may contribute to systematic errors in observation (Gräf & Unkelbach, 2016; Wirtz, 2017). Similarly, an expectancy effect may have influenced our evaluation of an observation. For instance, a former study found an expectancy effect regarding the rating of severity of current symptoms portrayed by SPs (Mumma, 2002), whereby the ratings were influenced by prior information about the SPs. Consequently, the SPs of the current study may have been detected had the trainees been given the information beforehand. On the other hand, in the study by Wünderich et al. (2012), it was noted that “in the majority of the cases, SPs were not detected” (p. 501). Nevertheless, several experimental studies do highlight the importance of controlling for and reducing rater expectancy and related effects (e.g., Ariel et al., 2019; Hoyt, 2002; Martell & Evans, 2005). Thus, it would be interesting – and important for future studies – to investigate whether psychotherapy trainees are more likely to accurately identify SPs if they were told about status differences beforehand compared to trainees who are ignorant of this fact.

Implications

In conclusion, SPs can potentially profit greatly regarding their authenticity by receiving thorough training. Thus, we strongly recommend training SPs before implementation. We conclude that, at an early stage of training, psychotherapy trainees are unable to distinguish between SPs and real patients, indicating that high levels of authenticity can be reached. Considering the new psychotherapy law, it would be of particular interest to replicate the current results with psychology students.

Electronic Supplementary Material

The electronic supplementary material is available with the online version of the article at <https://doi.org/10.1026/1616-3443/a000594>

ESM1. Rollenanleitung

References

- Adamo, G. (2003). Simulated and standardized patients in OSCEs: Achievements and challenges 1992–2003. *Medical Teacher*, 25(3), 262–270. <https://doi.org/10.1080/0142159031000100300>
- Anderson, S. F., & Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, 21(1), 1–12. <http://doi.org/10.1037/met0000051>
- Andrews, G., Henderson, S., & Hall, W. (2001). Prevalence, comorbidity, disability and service utilisation: Overview of the Australian National Mental Health Survey. *The British Journal of Psychiatry*, 178(2), 145–153.
- Angst, J., Gamma, A., Endrass, J., Goodwin, R., Ajdacic, V., Eich, D., & Rössler, W. (2004). Obsessive-compulsive severity spectrum in the community: Prevalence, comorbidity, and course. *European Archives of Psychiatry and Clinical Neuroscience*, 254(3), 156–164. <https://doi.org/10.1007/s00406-004-0459-4>
- Ariel, B., Sutherland, A., & Bland, M. (2019). The trick does not work if you have already seen the gorilla: How anticipatory effects contaminate pre-treatment measures in field experiments. *Journal of Experimental Criminology*, 55–66. <https://doi.org/10.1007/s11292-019-09399-6>
- Ay-Bryson, D. S., Weck, F., & Kühne, F. (in press). Can simulated patient encounters appear authentic? Development and pilot results of a rating instrument based on the portrayal of depressive patients. *Training and Education in Professional Psychology*. <https://doi.org/10.1037/tep0000349>
- Ay-Bryson, D. S., Weck, F., & Kühne, F. (under review). Can students in simulation portray a psychotherapy patient authentically with a detailed role-script? Results of a randomized-controlled study. *Manuscript submitted for publication*.
- Barrows, H. S., & Abrahamson, S. (1964). The programmed patient: A technique for appraising student performance in clinical neurology. *Journal for Medical Education*, 39(8), 802–805.
- Brem, B., Christen, R., Bauer, D., Wüst, S., Schnabel, K., & Woermann, U. (2018). *Handbuch für SchauspielpatientInnen der Medizinischen Fakultät der Universität Bern* [Handbook for actor patients at the medical faculty of the University of Bern]. Institut für Medizinische Lehre.
- Entwurf eines Gesetzes zur Reform der Psychotherapeutenausbildung, no. 19/9770 (2019). Retrieved from https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/3_Downloads/Gesetze_und_Verordnungen/GuV/P/Psychotherapeutenausbildung_Reform_Bundestag.pdf
- Gräf, M., & Unkelbach, C. (2016). Halo effects in trait assessment depend on information valence: Why being honest makes you industrious, but lying does not make you lazy. *Personality and Social Psychology Bulletin*, 42(3), 290–310. <https://doi.org/10.1177/0146167215627137>
- Hautzinger, M. (2013). *Kognitive Verhaltenstherapie bei Depressionen: Mit Online-Materialien* [Cognitive behavioral therapy of depression]. Beltz.

- Hodges, B. D., Hollenberg, E., McNaughton, N., Hanson, M. D., & Regehr, G. (2014). The psychiatry OSCE: A 20-year retrospective. *Academic Psychiatry*, 38(1), 26–34.
- Hoyt, W. T. (2002). Bias in participant ratings of psychotherapy process: An initial generalizability study. *Journal of Counseling Psychology*, 49(1), 35–46. <https://doi.org/10.1037/0022-0167.49.1.35>
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163.
- Krahn, L. E., Bostwick, J. M., Sutor, B., & Olsen, M. W. (2002). The challenge of empathy. *Academic Psychiatry*, 26, 26–30. <https://doi.org/10.1176/appi.ap.26.1.26>
- Kühne, F., Maaß, U., & Weck, F. (2020). Einsatz standardisierter Patienten im Psychologiestudium: Von der Forschung in die Praxis [Standardized patients in clinical psychology: From research to practice]. *Verhaltenstherapie*. <https://doi.org/10.1159/000509249>
- Kühne, F., Ay, D. S., Otterbeck, M., & Weck, F. (2018). Standardized patients in clinical psychology and psychotherapy: A scoping review on barriers and facilitators for implementation. *Academic Psychiatry*, 42(6), 773–781. <https://doi.org/10.1007/s40596-018-0886-6>
- Kühne, F., Heinze, P. E., & Weck, F. (2020). Standardized patients in psychotherapy training and clinical supervision: Study protocol for a randomized controlled trial. *Trials*, 21(276), 1–7. <https://doi.org/10.1186/s13063-020-4172-z>
- Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4), 355–362. <https://doi.org/10.1177/1948550617697177>
- Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, 1(3), 86–92.
- Martell, R. F., & Evans, D. P. (2005). Source-monitoring training: Toward reducing rater expectancy effects in behavioral measurement. *Journal of Applied Psychology*, 90(5), 956–963. <https://doi.org/10.1037/0021-9010.90.5.956>
- Melluish, S., Crossley, J., & Tweed, A. (2007). An evaluation of the use of simulated patient role-plays in the teaching and assessment of clinical consultation skills in clinical psychologists' training. *Psychology Learning & Teaching*, 6(2), 104–113.
- Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine*, 65(9), S63–S67. <https://doi.org/10.1097/00001888-199009000-00045>
- Mumma, G. H. (2002). Effects of three types of potentially biasing information on symptom severity judgments for major depressive episode. *Journal of Clinical Psychology*, 58(10), 1327–1345. <https://doi.org/10.1002/jclp.10046>
- Muse, K., & McManus, F. (2013). A systematic review of methods for assessing competence in cognitive-behavioural therapy. *Clinical Psychology Review*, 33(3), 484–499. <https://doi.org/10.1016/j.cpr.2013.01.010>
- Partschfeld, E. (2013). *Evaluation des Einsatzes von Simulationspatienten in der psychotherapeutischen Ausbildung* [Evaluation of the deployment of simulated patients in the psychotherapy training. A contribution to the empirical foundation of the psychotherapy training]: *Ein Beitrag zur empirischen Fundierung der Psychotherapieausbildung* [Doctoral dissertation].
- Partschfeld, E., Strauß, B., Geyer, M., & Philipp, S. (2013). Simulationspatienten in der Psychotherapieausbildung [Simulated patients in psychotherapy training. Evaluation of a teaching concept for development of therapeutic skills]. *Psychotherapeut*, 58(5), 438–445. <https://doi.org/10.1007/s00278-013-1002-8>
- Peters, T., & Thrien, C. (2018). *Simulationspatienten: Handbuch für die Aus- und Weiterbildung in medizinischen- und Gesundheitsberufen* [Simulated patients: Handbook for the training and advanced training in medical and health professions]. Hogrefe. <https://doi.org/10.1024/85756-000>.
- Scherer, M., & Ehrhardt, M. (2017). *Handbuch für Simulationspatienten am Universitätsklinikum Hamburg-Eppendorf*. Institut für Allgemeinmedizin.
- Sharpless, B. A., & Barber, J. P. (2009). A conceptual and empirical review of the meaning, measurement, development, and teaching of intervention competence in clinical psychology. *Clinical Psychology Review*, 29(1), 47–56.
- Sheen, J., Sutherland-Smith, W., Thompson, E., Youssef, G. J., Dudley, A., King, R., Hall, K., Dowling, N., Gurtman, C., & McGillivray, J. A. (2020). Evaluating the impact of simulation-based education on clinical psychology students' confidence and clinical competence. *Clinical Psychologist*. <https://doi.org/10.1080/13284207.2021.1923125>
- Voderholzer, U. (Hrsg.). (2007). *Lehre im Fach Psychiatrie und Psychotherapie: Ein Handbuch* [Teaching in the subject psychiatry and psychotherapy: A handbook]. Kohlhammer.
- Wirtz, M. A. (2017). Interrater Reliability. In *Encyclopedia of Personality and Individual Differences* (pp. 1–4). Springer International Publishing AG. https://doi.org/10.1007/978-3-319-28099-8_1317-1
- Wünderich, M. S., Nissen, C., Feige, B., Philipsen, A. S., & Voderholzer, U. (2012). Portrayal of psychiatric disorders: Are simulated patients authentic? *Academic Psychiatry*, 36(6), 501–502. <https://doi.org/10.1176/appi.ap.11090163>

Acknowledgments

We would like to thank all psychotherapy trainees, SPs and patients for their participation in the study as well as Samuel Oliver Ay-Bryson for proofreading the manuscript.

Conflict of interest

The authors declare no conflicts of interest.

Funding

The Hans Böckler Foundation supported this research with a full PhD scholarship to Destina Sevde Ay-Bryson. Open Access-Publication enabled by University of Potsdam.

Destina Sevde Ay-Bryson

Clinical Psychology and Psychotherapy
University of Potsdam
Karl-Liebknecht-Str. 24–25
14476 Potsdam
Germany
destina.ay@uni-potsdam.de