Research in Science Education (2023) 53:139–153
https://doi.org/10.1007/s11165-022-10046-w

Assessing Pre-service Teachers' Views of Scientists, Their Activities, and Locations: the VoSAL Instrument

Bianca Reinisch[1] · **Moritz Krell**[2]

Accepted: 16 January 2022 / Published online: 15 February 2022

Abstract
In science education, learners' conceptions of scientists and their work are often assessed by the Draw-A-Scientist Test (DAST). Due to validity concerns, methodical literature demands the development of alternative instruments to measure learners' conceptions validly and efficiently. This study presents an instrument with 29 rating scale items to assess pre-service teachers' (PSTs) Views of Scientists, their Activities, and Locations (VoSAL). The items were developed based on theoretical considerations, previous findings, and repeated discussions by biology education experts. After several steps of test development, PSTs filled out the questionnaire ($N = 1{,}098$). Exploratory factor analyses and reliability measurements mostly confirm the proposed structure. Groups comparisons were performed regarding the results from pre-service biology teachers of three different study stages ($n_{freshmen} = 114$; $n_{second\ and\ third\ years} = 124$; $n_{graduates} = 107$). Analyses of variance and corresponding post-hoc tests showed that undergraduates (freshmen, second and third years) differ significantly from graduates regarding the scales *stereotypical appearance*, *inquiry location*, and *scientific activity*, with undergraduates having more stereotypical conceptions than graduates. In sum, the VoSAL can be utilized to gain valid data of PSTs' conceptions about scientists and their work. Also, the VoSAL can be considered efficient since the test time is between 5 and 10 min. Thus, the questionnaire is valuable in studies that aim to introduce and expose PSTs to realistic science images.

Keywords Questionnaire · Nature of science · Scientists · PSTs · Scientific activities · Validity

Introduction

Science education research has shown that students enter science classes with conceptions of phenomena that often differ from scientific views (Treagust & Duit, 2008), including conceptions about scientists and their work (DeWitt et al., 2013; Finson, 2002). The latter

✉ Bianca Reinisch
bianca.reinisch@fu-berlin.de

[1] Biology Education, Freie Universität Berlin, Schwendenerstr. 1, 14195 Berlin, Germany

[2] Department Biology Education, IPN – Leibniz Institute for Science and Mathematics Education, Olshausenstr. 62, 24118 Kiel, Germany

is part of a nature of science understanding (McComas & Clough, 2020), which is vital for students' and teachers' scientific literacy (Roberts, 2007).

There are strong indications that teachers' attitudes toward science and their images about science and scientists influence their students' views (Christidou, 2011; van Tuijl & van der Molen, 2016). At the same time, teachers "may suffer from lack of knowledge on work and workers in [the] fields" (van Tuijl & van der Molen, 2016, p. 173). Consequently, contents about scientists and their work need to be considered in teacher education.

The stereotypical students' and pre-service teachers (PSTs)' view of a scientist mainly relates to the appearance of a scientist and includes a male, elderly or middle-aged, lab coat wearing, bespectacled person in a laboratory doing (dangerous) experiments on his own (DeWitt et al., 2013; Finson, 2002). One of the most commonly used instruments to assess individuals' conceptions of scientists is the Draw-A-Scientist Test (DAST; e.g., Finson et al., 1995) or modified versions of it (e.g., Farland, 2006). However, several studies criticized that these instruments do not provide a valid assessment of individuals' conceptions (e.g., Losh et al., 2008; Reinisch et al., 2017). In response to these criticisms, the present study aims to develop an instrument to assess PSTs' conceptions of scientists and their work validly and efficiently.

Theoretical Background

Nature of Science and Its Relevance in Teacher Education

The acquisition of knowledge about the nature and methods of science, the appreciation of its history and development, and the awareness of the manifold connections between science, technology, society, and environment is deemed to be essential to achieve scientific literacy for students and (pre-service) teachers (Hodson, 2014). Corresponding research focuses on questions such as "What is science?," "How does science work?," "How does science impact and is impacted by society?," and "What are scientists like in their professional and personal lives?" (McComas & Clough, 2020, p. 5). These questions should also be discussed in the science classroom (e.g., KMK, 2005, 2020; NGSS Lead States, 2013) and, hence, in science teacher education courses (e.g., KMK, 2019, 2020; NSTA & ASTE, 2020).

Against this background, the question arises: What images of scientists and their work do students and (pre-service) teachers typically have? Hence, developing and evaluating instruments suitable to assess students' and (pre-service) teachers' views about scientists and their work is one crucial part of research in science education (Chang et al., 2020; Reinisch et al., 2017). In the recent research literature, such instruments typically address three aspects: the appearance of scientists, the activities scientists perform, and the location scientists work at (Farland-Smith, 2017; Lamminpää et al., 2020; Reinisch et al., 2017).

Views of Scientists, Their Activities, and Locations

Related to the appearance of a scientist, most DAST studies revealed a standard image with a male scientist, who has a lab coat, eyeglasses, and facial hair (Chambers, 1983; Finson, 2002). In recent studies, male and female scientists were drawn by the respondents more equally (Miller et al., 2018), although other stereotypical features of scientists such as lab coats and eyeglasses remain prominent (Lamminpää et al., 2020; Reinisch et al., 2017).

Although the traditional DAST does not explicitly ask for drawing scientific activities, some respondents still address them in their drawings. The most prominent activity found in such studies refers to laboratory, mostly experimental work, sometimes accompanied by dangerous explosions (Chambers, 1983; Finson, 2002). In recent DAST studies, a modified prompt explicitly asking for scientific activities is given to the respondents (Farland-Smith, 2017; Lamminpää et al., 2020; Reinisch et al., 2017). In these studies, inquiry activities (e.g., experimenting, evaluating, thinking) were mostly detected among the received drawings (Lamminpää et al., 2020; Reinisch et al., 2017).

In their interview study with scientists, Wentorf et al. (2015) used the RIASEC model, which is named after six professional activities (Realistic, Investigative, Artistic, Social, Enterprising, Conventional) postulated by Holland (1963). Wentorf et al. (2015) found a seventh category, which they called networking (RIASEC + N; Table 1). On this basis, they developed the Nature of Scientists questionnaire to analyze students' views about scientists. Overall, the 14–15-year-old students surveyed mostly agreed to realistic, investigative, and artistic activities. While the students attached greater importance to networking activities, enterprising activities were rated less important.

Concerning the location scientists work at, typically children but also adults show a laboratory setting, in which symbols of research (e.g., laboratory equipment), symbols of knowledge (e.g., books), and technology items (e.g., computers) are included (Chambers, 1983; Finson, 2002; Reinisch et al., 2017). For example, the Draw-A-Science Comic instrument by Lamminpää et al. (2020) revealed comics that mostly showed laboratories, classrooms, or outdoor locations.

Methodological Challenges Concerning the Draw-A-Scientist Test

The DAST has become one of the most established instruments to assess students' conceptions about scientists. The existing studies using the DAST or modified versions such as the Draw-A-Science Comic (Lamminpää et al., 2020) provided valuable insights into this domain, which informed the development of teaching approaches and posed avenues for further research (Chang et al., 2020; Finson, 2002).

Despite many advantages of the DAST (Chambers, 1983; Farland, 2006; Finson, 2002) and drawing assessments in general (Chang et al., 2020; Finson & Pederson, 2011), there are some challenges regarding the analysis and interpretation of the drawings. Detailed discussions of such issues can be found elsewhere (e.g., Lamminpää et al., 2020; Losh et al., 2008; Reinisch et al., 2017). Two critical points can be noted here as examples: First,

Table 1 Characteristics and exemplary activities of scientists based on the RIASEC + N model (Wentorf et al., 2015, p. 213)

Category	Characterization	Example
Realistic	Skilled manual activities	Manufacture substances in the laboratory
Investigative	Analytical, intellectual activities	Evaluate results from experiments
Artistic	Creative activities	Develop ideas for new research approaches
Social	Caring, interpersonal activities	Carry out a course
Enterprising	Entrepreneurial activities	Raise funds for research projects
Conventional	Precise, structured activities	Do administrative tasks
Networking	Collaborating activities	Exchange with scientists from other groups

studies show that the graphic abilities of the respondents influence the outcome of the drawings (Losh et al., 2008; Reinisch et al., 2017) Second, it is questionable if a drawing really represents someone's conception or rather shows the adoption of a common representation (Finson & Pederson, 2011; Reinisch et al., 2017).

Considering methodological criticism regarding the valid interpretation of drawings, alternative assessment formats to overcome this issue need to be developed (Lamminpää et al., 2020; Reinisch et al., 2017). In other areas of nature of science research, closed-ended instruments proofed to be a valid alternative to assess learners' conceptions (e.g., Edgerly et al., 2021; Liang et al., 2008; Urhahne et al., 2011). For example, Edgerly et al. (2021) used the Students' Understanding of Science and Scientific Inquiry (SUSSI) instrument (Liang et al., 2008) to assess elementary teachers NOS understanding by Likert scale items. They found that four of the eight NOS constructs had acceptable Cronbach's Alpha levels. The authors conclude "that quantitative assessment could detect interesting differences in participants' NOS views" (Edgerly et al., 2021, p. 1).

Aim and Research Questions

This study aims to develop an instrument to assess PSTs' conceptions of scientists and their work validly and efficiently. The main research question for this study was: *To what extent is it possible to validly assess PSTs' conceptions about scientists, scientific activities, and scientific locations in an efficient way (i.e., using closed-ended questions)*?

In psychological and educational research, validity is understood as a process of creating and evaluating arguments, which provide a solid scientific basis for the interpretation and use of test scores (argument-based approach to validation; Kane, 2013). In this study, we aim to develop and analyze an instrument considering validity evidence based on internal structure and based on relations to other variables (American Educational Research Association (AERA) et al., 2014).

Validity evidence based on internal structure refers to "the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (AERA et al., 2014, p. 16). Our first research question (RQ) is as follows:

To what extent does data support the intended structure of the instrument, representing the scales *appearance* of scientists, the *activities* scientists perform, the *location* scientists work at, and related subscales?

We expect that items, which have been developed to assess the same scale, are reflected as such in factor analyses. Furthermore, the internal consistency of each scale is expected to be sufficiently high (i.e., reliability values of Cronbach's alpha (α) > 0.70; Field, 2018).

Validity evidence based on relations to other variables relies on the comparison of the test with external variables. Known groups comparisons are a possibility to evaluate whether an instrument can discriminate between groups known to differ on the variable of interest (AERA et al., 2014). The second research question is as follows:

To what extent is the instrument able to elicit conceptions about the *appearance* of scientists, the activities scientists perform, and the *location* scientists work at from different samples of PSTs in an expected way?

It can be assumed that the instrument detects differences between different groups of PSTs. We expected that graduates among PSTs have less stereotypical views about scientists and their work than undergraduates. Meta studies show that most conceptions

regarding scientists and their work remain stereotypical — regardless of when and with whom the assessment was taken (Finson, 2002). However, there are some significant shifts in terms of certain elements (e.g., scientists being male, presence of a lab coat) being less present with increasing age of the study participants (Miele, 2014; Milford & Tippett, 2013; Miller et al., 2018). More generally, a recent study suggests that the nature of science understanding of pre-service biology teachers (PSBTs) increases throughout teacher education at university (Bruckermann et al., 2018).

Study Design: Addressing Sources of Validity Evidence

The development and evaluation of the instrument, which is named the views of scientists, activities, and locations instrument (VoSAL), followed three sequential steps: (1) the initial drafting of the instrument, (2) a pilot study with subsequent item revisions, and (3) the main study to evaluate the instrument. During the development of the VoSAL, the number of items was altered due to the adaption of the instrument considering the results of the preceding studies (Table 2). For both the pilot and the main study, PSTs of one university in Germany participated voluntarily and anonymously. In Germany, PSTs typically study two subjects (i.e., two future teaching subjects). In the following, the three steps will be further delineated, including the consideration of addressing sources of validity evidence.

Initial Drafting of the Instrument and Pilot Study

The VoSAL has been developed strongly based on theoretical considerations and previous findings to align to content validity (AERA et al., 2014). Five point-rating scale items ("not true" to "totally true") for the three scales *appearance* of scientists, *activities* scientists perform, and *location* scientists work at were constructed. The category system presented by Reinisch et al. (2017) was used to develop initial items for the scales *appearance* and *location*. For the scale *activity*, the Nature of Scientists questionnaire (Wentorf et al., 2015) was used and slightly adapted for the sample of PSTs. The first draft of the instrument was

Table 2 Number of items in the different study steps; for example items, see below

Scale	The number of items after the …		
Subscale	… initial drafting	… pilot study	… main study
Appearance	12	5	5
Location	6	7	4
Activity	5	3	3
Realistic	6	3	3
Investigative	5	3	3
Artistic engineering	4	3	3
Social	8	6	3
Enterprising	3	3	3
Conventional	3	3	2
Networking			
Σ	52	36[a]	29[a]

[a]Number of items included after item selection of the respective study step

individually reviewed by experts in the field of biology education ($N = 5$) and subsequently discussed with all of them. We administered the resulting 52-item instrument in a pilot study among 92 PSBTs (undergraduates). Exploratory factor analyses (EFA) and reliability analyses were performed using the software IBM SPSS Statistics to test for the internal structure of the instrument. Based on the results and content-related considerations, the instrument was revised and then contained 36 items (Table 2).

Evaluation of the Instrument: Main Study

For the main study, 1,098 PSTs were requested to answer the VoSAL. To address RQ1, EFAs and reliability analyses were performed again. For EFAs, principal component analysis with varimax rotation was used and the Kaiser criterion (eigenvalue greater than 1) was applied to determine the number of extracted factors (Field, 2018).

To address RQ2, data from PSBTs in three different study stages were compared (known groups comparisons; AERA et al., 2014): 114 freshmen, who were at the beginning of their first semester (cohort BA_1), 124 second and third years (cohort BA_2), and 107 graduate students enrolled in a Master of Education program (cohort MA).

Analyses of variance (ANOVAs) were conducted to check for differences between the three cohorts, with corresponding post-hoc tests (Bonferroni) to compare the single cohorts. For the interpretation of the effect size measure eta squared (η^2), Cohen's (1988) recommendation of small (> 0.01), medium (> 0.06), and large (> 0.14) effects was applied (Fritz et al., 2012). If the assumption of homogeneity of variances was violated, Welch's F-ratios were considered and the Games-Howell post-hoc test was applied (Field, 2018).

Results

In the following, results from the main study will be reported.

Results for the Internal Structure of the Instrument (RQ 1)

Separate EFAs were performed for the items of each of the three scales *appearance*, *location*, and *activity*. For all EFAs, the Kaiser–Meyer–Olkin (KMO) measures verified the sampling adequacy (KMO > 0.60; "mediocre"; Kaiser, 1974) and all KMO values for individual items were above the acceptable limit of 0.5 (Field, 2018). Also, Bartlett's test of sphericity indicated that correlations between items were sufficiently large ($p < 0.001$).

The EFA with the five items of the scale *appearance* revealed one factor with an eigenvalue greater than 1, which in sum explained 50.10% of the variance (Table 3). The five items indicate a stereotypical conception of scientists wearing lab coats, safety glasses, etc.

For the scale *location*, the EFA revealed two factors with eigenvalues greater than 1, which in sum explained 55.19% of the variance. Table 4 shows the factor loadings after rotation. Factor one includes items that refer to places outdoor and in the lab. Comparing these items with items of factor 2, there might be a connection to the kind of activities performed there. That is, places represented in items of factor 1 might be connected to inquiry activities, while items of factor 2 can be connected to a broader range of activities.

The instrument contains 20 items for the scale *activity* (Table 2). The EFA revealed five factors with eigenvalues greater than 1, which in sum explained 56.34% of the variance.

Table 3 Result of exploratory factor analysis for the scale *appearance* ($N = 1{,}098$)

Item	Label	Factor 1
The typical natural scientist …		
… wears a lab coat at work	App_11	.841
… wears protective cloth at work	App_12	.833
… wears safety glasses at work	App_09	.800
… wears everyday clothes at work	App_10	.649
… is rather old (older than 50 years)	App_05	.203
Eigenvalue		2.51
% of variance		50.10

The original version of the VoSAL is in the German language and linguistic flaws may be caused by the translation

Table 4 Result of exploratory factor analysis for the scale *location* ($N = 1{,}098$)

Item	Label	Factor 1	Factor 2
The typical natural scientist works …			
… in the nature	Loc_4	.836	− .240
… outside	Loc_6	.795	− .303
… in the open country	Loc_5	.608	
… in the lab	Loc_1	.605	.369
… inside	Loc_3		.793
… in the office	Loc_2		.583
… at university	Loc_7	.442	.566
Eigenvalue		2.28	1.58
% of variance (after rotation)		32.58	22.61

Factor loadings < .2 are not shown; the original version of the VoSAL is in the German language and linguistic flaws may be caused by the translation

Table 5 shows the factor loadings after rotation. Items of the assumed subscales *investigative* and *realistic* loaded on the same factor; items of the assumed subscales *enterprising* and *networking* also loaded on one factor with one *networking*-item (Net_3) being deleted due to the main loading on factor 1 (not shown in Table 5).

Several reliability analyses were performed for the scales *appearance*, *location*, *activity*, and corresponding subscales as found in the EFAs. Results of these analyses for the items, which have been selected for further analysis, are shown in Table 6. As the final items of the scale *appearance* collectively refer to a rather *stereotypical appearance* of scientists, this scale is renamed accordingly. For *location*, items of factor 2 (Table 4) have a relatively low reliability ($\alpha = 0.42$; Field, 2018). Hence, these items were not used for further analysis. Items of factor 1 ($\alpha = 0.70$) are subsumed under the scale *inquiry locations* (Table 6) as all items refer to places in which mostly inquiry activities can be performed.

Contrary to the results of the EFA for the scale *activity* (Table 5), a separate reliability analysis was performed for each of the *investigative* and the *realistic* subscale. Both Cronbach's α values are sufficiently high considering the small number of items (Table 6). Also, the content of the two subscales is separable. Likewise, separate reliability analyses were performed for the subscales *enterprising* and *networking* as found by Wentorf et al. (2015).

Table 5 Result of exploratory factor analysis for the scale *activity* ($N = 1{,}098$)

Item	Label	F 1	F 2	F 3	F 4	F 5
The typical natural scientist performs the following activities regularly:						
Analyze and interpret results from experiments	Inv_2	.752				
Recognize relationships in measured data	Inv_3	.750				
Develop ideas for new research approaches	Inv_5	.435	.273		.478	
Make a protocol	Rea_2	.728				
Carry out an investigation	Rea_3	.726				
Perform measurements	Rea_1	.720				
Build and manage a team	Ent_5		.720			
Lead a research group	Ent_2		.683			
Organize and lead projects	Ent_1		.572			
Hold meetings with colleagues from other departments	Net_2	.202	.601			
Carry out interdisciplinary projects	Net_1	.229	.518			
Plan and manage finances	Con_1			.833		
Do administrative tasks	Con_2			.776		
Complete accounts for research funds used	Con_3			.756		
Work on inventions	Art_4				.750	
Develop measurement methods	Art_2	.210			.740	
Construct experimental equipment	Art_3	.287			.645	
Accompany students' theses	Soc_3			.208		.777
Supervise students	Soc_1		.239			.764
Prepare and conduct courses	Soc_2					.728
Eigenvalue		4.53	2.74	1.59	1.28	1.14
% of variance (after rotation)		15.84	11.26	10.22	9.90	9.13

Factor (F) loadings < .2 are not shown; item labels refer to the original seven subscales (RIASEC + N) as found by Wentorf et al. (2015); the original version of the VoSAL is in the German language and linguistic flaws may be caused by the translation

Table 6 Results of reliability analyses (Cronbach's α) for all (sub)scales ($N = 1{,}098$)

Scale	n_{items}	M	SD	α
Stereotypical appearance	5	3.29	0.67	.70
Inquiry locations	4	3.47	0.65	.70
Activity				
Realistic	3	4.18	0.65	.70
Investigative	3	4.24	0.60	.68
Artistic	3	3.40	0.75	.63
Teaching	3	3.09	0.68	.70
Conventional	3	2.76	0.80	.73
Social	5	3.44	0.59	.66

Compared to Table 2, n_{items} are different for the *activity* subscales as some of them were merged

However, Cronbach's α was relatively low ($\alpha = 0.63$ and 0.48, respectively). Considering that the corresponding items also loaded on one common factor (Table 5), one reliability analysis was performed, which showed a sufficiently high Cronbach's α (Table 6). This finding also matches that all items present social activities within a research group (previous *enterprising* items) and outside with other researchers (previous *networking* items). Hence, the items are subsumed under the *social* subscale. The items by Wentorf et al. (2015) previously subsumed under the *social* scale are now renamed into *teaching* to consider the content of the corresponding items more strongly (e.g., "supervise students").

Considering the previous explanations, Cronbach's α values are sufficiently high for all (sub)scales considering the relatively small number of items (Field, 2018).

Results for the Known Groups Comparisons (RQ 2)

With the selected items (Table 6), groups comparisons were conducted to address RQ 2. It was expected that graduates have less stereotypical views about scientists and their work than undergraduates (e.g., Bruckermann et al., 2018). For the scales *stereotypical appearance* and *inquiry locations*, there is a significant effect of the cohort on the PSBTs' scores, with small effect size measures. For both scales, significant differences between BA_1 and MA were found in post-hoc tests (Table 7), with graduate students (MA) having less stereotypical views about scientists and their working place than freshmen (BA_1; Table 8).

Table 7 Results of ANOVAs, comparing the cohorts BA_1, BA_2, and MA for all (sub)scales

Scale			p	η^2	Post-hoc test		
	df	F			BA_1–BA_2	BA_1–MA	BA_2–MA
Stereotypical appearance[1]	2,226.957	10.680	< .001	.052	.289	.000	.017
Inquiry locations	2,341	8.695	< .001	.049	.012	.000	.590
Activity							
Realistic	2,341	6.504	.002	.037	1.000	.011	.003
Investigative	2,342	3.318	.037	.019	.857	.033	.349
Artistic	2,342	3.996	.019	.023	.085	.026	1.000
Teaching	2,342	2.613	.075	.015	.878	.664	.069
Conventional	2,342	14.446	< .001	.078	1.000	.000	.000
Social[1]	2,342	1.501	.177	.009	.625	.150	.692

Effect size eta squared: $\eta^2 > .01$ small effect, $\eta^2 > .06$ medium effect (Fritz et al., 2012), Bonferroni post-hoc test; [1]based on Welch's F-ratio and Games-Howell post-hoc test (Field, 2018)

Table 8 Mean scores and standard deviations for the cohorts BA_1, BA_2, and MA for the scales *stereotypical appearance* and *inquiry locations*

Scale	Cohort	N	M	SD
Stereotypical appearance	BA_1	114	3.53	.61
	BA_2	124	3.39	.75
	MA	107	3.15	.64
Inquiry locations	BA_1	114	3.73	.63
	BA_2	124	3.49	.67
	MA	106	3.37	.69

Based on the literature (Finson, 2002; Lamminpää et al., 2020; Reinisch et al., 2017; Wentorf et al., 2015), a rather stereotypical view of scientific *activities* would include a high agreement to *realistic* and *investigative* activities and a relatively low agreement to other activities. In Fig. 1, the mean scores and corresponding standard deviations (SD) for the three cohorts for each *activity* scale are depicted. Overall, the PSTs more strongly agreed to the *realistic* and *investigative* activities than to the *artistic*, *teaching*, *conventional*, and *social* activities of scientists. Regarding a comparison of the three cohorts, post-hoc tests for each subscale showed significant differences for all subscales except *teaching* and *social* with mostly small effect sizes (Table 7). Also, post-hoc tests revealed significant differences between freshmen (BA_1) and graduates (MA) for four subscales and between second and third years (BA_2) and graduates (MA) for two subscales (Table 7).

Next, it can be assumed that a less stereotypical view of scientific *activities* would reveal a more balanced image of scientists performing a wide variety of activities. In Fig. 1, the shade for group MA is the narrowest, and hence, this group agreed more equally to all activities than the other two groups. To evaluate this visual impression, the mean score and corresponding SD across all items were calculated. We assumed that the lower the SD (as a measure of spread), the more consistently the PSTs responded and, in this case, the more equally they agreed to all activities. A Welch ANOVA was conducted to test for differences in the calculated SDs between the three cohorts. Results revealed that the SDs in the responses of the three groups differ significantly (Table 9). Games-Howell post-hoc tests showed a significant difference between the groups BA_1 and MA and between BA_2

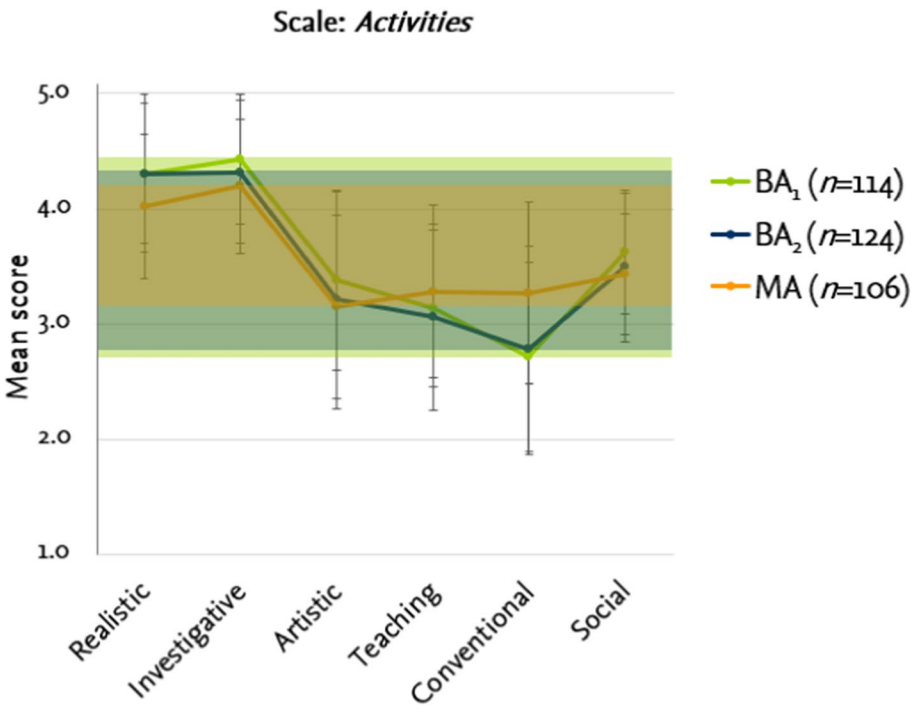

Fig. 1 Mean scores and standard deviations for activities across the cohorts (BA_1, BA_2, MA); the three shades illustrate the range of the mean scores of the three cohorts

Table 9 Results of Welch ANOVA, comparing SD across all items between the cohorts BA_1, BA_2, and MA

				Post-hoc test		
	df	F	*p*	BA_1–BA_2	BA_1–MA	BA_2–MA
SD across all items	2,227.294	12.902	< .001	.347	.002	.000

Games-Howell post-hoc test (Field, 2018)

and MA but not between BA_1 and BA_2. These findings indicate that responses across all activity items are statistically more equal to each other for graduates than for second and third years.

Discussion

This study aimed to develop an instrument to assess PSTs' conceptions of scientists and their work validly and efficiently. For this, different sources of validity evidence (AERA et al., 2014; Kane, 2013) were addressed, which will be evaluated in the following.

Discussion of the Internal Structure of the Instrument (RQ 1)

The multistep procedure in this study led to the selection of items, which are subsumed under eight (sub)scales. Both Bartlett's tests and the KMO measures of sampling adequacy indicate that the variables are suitable for the factor analyses. Results of the reliability analyses can be rated as satisfactory for most of the scales ($0.63 \leq \alpha \leq 0.73$; Field, 2018).

Two lines of research reason the initial construction of the items: First, theoretical considerations and results of DAST studies (Finson, 2002) were used to construct items for the appearance of scientists and the location scientists work at (Reinisch et al., 2017). In corresponding studies, received drawings from participants are usually evaluated regarding underlying stereotypical conceptions of scientists (e.g., a scientist wearing a lab coat; Finson, 2002). An EFA showed that items, which correspond to this stereotypical conception, are also loading on one factor (Table 3) and, hence, remain as one scale after the final item selection (*stereotypical appearance*).

Drawings of scientific locations which are typically regarded as stereotypical contain laboratory settings (Finson, 2002). In recent studies, additional locations were drawn, such as classrooms or outdoor locations (Lamminpää et al., 2020). For the initial item construction, such locations were considered. An EFA revealed that items that first do not seem to belong together (i.e., laboratory and outdoor locations; Table 4) load on the same factor. These items primarily reflect locations in which inquiry activities can be performed. A follow-up study with more *location* items could reveal to what extent certain locations relate to specific activities in students' conceptions. Lamminpää et al. (2020) emphasized the meaning of students' conceptions about scientific activities, which were depicted the most frequently in the drawn comics of the assessed students. Hence, it is reasonable to assume an influence of the scale *activities* on the scale *locations*.

The second line of research, which was used to construct the VoSAL, refers to the RIASEC + N model. In particular, items of the Nature of Scientists questionnaire (Wentorf et al., 2015) were used to assess PSTs' conceptions of scientific activities. These items

offer a range of potential activities scientists perform (Table 1). Reinisch et al. (2017) used the underlying separation into seven activities (RIASEC + N) as a basis to analyze DAST drawings from PSBTs. They reported on the difficulty of separating between drawn investigative and realistic activities. Wentorf et al. (2015), who assessed students' conceptions, found seven scales using confirmatory factor analysis. However, they also report that confirmatory factor analysis of the data received from students' answers to their interest in scientific activities and their self-efficacy regarding these activities revealed only five scales. That is, items for interest in and self-efficiency of realistic, investigative, and artistic activities were subsumed under the scale inquiry. Our EFA revealed five factors with realistic and investigative items loading on the same scale. Also, item Inv_5 loaded on two other factors, with one cross-loading being significant (>.3; Table 5; Hair et al., 2014). Although not significant, there are also cross-loadings of the items Art_2 and Art_3 on factor 1 (*investigative* scale; Table 5). In this study, three separate scales were assumed for further analysis due to satisfactory reliability values (Table 6) and content-related considerations. Nevertheless, the findings of the EFA hint at the fact that further study should be done as to what extent the three scales *realistic*, *investigative*, and *artistic* can be regarded as separate scales, for example, by refining existing or developing new items.

Contrary to the original RIASEC-construct by Holland (1963), Wentorf et al. (2015) found a seven-scale construct for students' conceptions (RIASEC + N). That is, they additionally identified the scale networking. However, they emphasize the need for further studies as confirmatory factor analysis was ambiguous for the existence of this scale. In the present study, an EFA showed that two of the three original networking items loaded on the same factor as the items for the previously assumed scale *enterprising* (F 2; Table 5). The content of all items reveals activities that are strongly aligned to the social exchange with other scientists — either within the own research group or with scientists working at different places of work. Hence, the factor is named *social*. Differences to the identified structure of the RIASEC + N model by Wentorf et al. (2015) might be explained by the different sample assessed in this study. A primary difference is that PSTs study at university and are closer to scientists than school students. However, it remains unclear and would need more investigation as to what extent this affects their views.

Discussion of the Known Groups Comparisons (RQ 2)

With the conducted groups comparisons, this study aimed to gain validity evidence for the VoSAL by comparing the PSTs' answers with external variables. We expected that graduates have less stereotypical views about scientists and their work than undergraduates (e.g., Bruckermann et al., 2018; Miele, 2014). Results showed significant differences with small to medium effect sizes between the three groups (BA_1, BA_2, MA) for most of the (sub) scales (Table 7).

PSBTs' answers for the *stereotypical appearance* scale showed differences in dependence of them being Bachelor or Master students. Master students have significantly fewer stereotypical conceptions about the appearance of scientists than Bachelor students. As discussed earlier, the scale *inquiry location* reflects locations that can be connected foremost to inquiry activities. Such activities are typically regarded as stereotypical (Chambers, 1983; Finson, 2002; Reinisch et al., 2017), and thus, a high agreement to items of the scale *inquiry location* might reflect a rather stereotypical conception. Results revealed a decrease of agreement along study stages (Table 8) and hence, a less stereotypical conception in later stages of study. Findings for both scales *stereotypical appearance* and *inquiry*

location align with a more general notion of an adequate NOS understanding over the course of studies (Bruckermann et al., 2018). The difference between Bachelor and Master students might be explained by the fact that all Bachelor students complete a Bachelor thesis before entering the Master study. Typically, they are then supervised by a practicing scientist in authentic working environments. However, not every PSBT necessarily conducts a bachelor thesis in the field of biology but might conduct it in their other subject (e.g., German, History) as well. Consequently, it would be interesting to investigate possible differences between different groups of PSTs depending on their practical experience in a scientific field. Working together with practicing scientists might shape conceptions of scientists and their working places in general.

Considering the results for the subscales of the *activities* scale, the expectation was met in terms of a higher agreement to realistic and investigative activities by the PSBTs still in their bachelor studies (BA_1, BA_2) compared to those in their master studies (MA; Fig. 1). A lower approval rate to inquiry activities (i.e., *realistic*, *investigative*) from the PSTs that are further progressed in their studies (MA) seems to go along a higher approval rate to other less stereotypical activities. This is particularly evident for conventional activities (Fig. 1). Such activities include, for example, administrative activities that are unlikely to be taught much or not at all in science courses. PSTs can only get impressions of this when they perceive a regular scientific job. More generally, it can also be seen that Master students agree to the various activities in a more balanced manner than Bachelor students (Table 9). Accordingly, it can also be assumed that specific insights into the scientific environment of the university, such as when writing a Bachelor thesis, broaden the view of scientific activities.

In sum, the VoSAL seems to detect differences between different groups of PSBTs regarding their conceptions about scientists and their work with consideration of the study stage.

Conclusion and Prospects

This study addressed validity evidence based on internal structure and relations to other variables (i.e., known groups comparisons; AERA et al., 2014) for the VoSAL. For both, validity evidence sources arguments were provided.

We recommend investigating additional sources of validity evidence, for example, based on response processes by conducting a think-aloud study (AERA et al., 2014; Hubley & Zumbo, 2017).

We propose that the VoSAL can serve as an alternative to the widely known DAST as, for example, no drawing skills are needed (Losh et al., 2008; Reinisch et al., 2017). As a next step, however, a direct comparison of both instruments could be fruitful for further insights (Chang et al., 2020).

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflicts of Interest The second author, Moritz Krell, was employed at the same university (Freie Universität Berlin, Biology Education) as the first author, Bianca Reinisch, during the time the study was conducted. Other than that, the authors have no conflicts of interest to declare that are relevant to the content of this article.

References

Abd-el-Khalick, F. (2013). Teaching with and about nature of science, and science teacher knowledge domains. *Science & Education, 22*, 2087–2107.

AERA (American Educational Research Association), American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington D.C.: AERA.

ASTA (Australian Science Teacher Association). (2009). *National professional standards for highly accomplished teachers of science: Final draft*. ASTA.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., & Tsai, Y.-M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal, 47*, 133–180.

Bruckermann, T., Ochsen, F., & Mahler, D. (2018). Learning opportunities in biology teacher education contribute to understanding of nature of science. *Education Sciences*, *8*(3), 1–18.

Chambers, D. W. (1983). Stereotypic images of the scientist. *Science Education, 67*, 255–265.

Chang, H.-Y., Lin, T.-J., Lee, M.-H., Lee, S.W.-Y., Lin, T.-C., Tan, A.-L., & Tsai, C.-C. (2020). A systematic review of trends and findings in research employing drawing assessment in science education. *Studies in Science Education, 56*(1), 77–110.

Christidou, V. (2011). Interest, attitudes and images related to science. *International Journal of Environmental & Science Education, 6*, 141–159.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Erlbaum.

DeWitt, J., Archer, L., & Osborne, J. (2013). Nerdy, brainy and normal: Children's and parents' constructions of those who are highly engaged with science. *Research in Science Education, 43*, 1455–1476.

Dove, J. E., Everett, L. A., & Preece, P. F. W. (1999). Exploring a hydrological concept through children's drawings. *International Journal of Science Education, 21*, 485–497.

Edgerly, H. S., Kruse, J. W. & Wilcox, J. L. (2021). Quantitatively investigating inservice elementary teachers' nature of science views. *Research in Science Education* . https://doi.org/10.1007/s11165-021-09993-7.

El Takach, S., & Yacoubian, H. A. (2020). Science teachers' and their students' perceptions of science and scientists. *International Journal of Education in Mathematics, Science and Technology, 8*(1), 65–75.

Farland, D. (2006). The effect of historical, nonfiction trade books on elementary students' perceptions of scientists. *Journal of Elementary Science Education, 18*(2), 31–47.

Farland-Smith, D. (2017). The evolution of the analysis of the draw-a-scientist test. In P. Katz (Ed.), *Drawing for Science Education*. Rotterdam: BRILL.

Field, A. (2018). *Discovering statistics using IBM SPSS statistics*. Sage.

Finson, K. D. (2002). Drawing a scientist. *School Science and Mathematics, 102*, 335–345.

Finson, K., & Pederson, J. (2011). What are visual data and what utility do they have in science education? *Journal of Visual Literacy, 30*, 66–85.

Finson, K. D., Beaver, J. B., & Cramond, B. L. (1995). Development and field test of a checklist for the draw-a-scientist test. *School Science and Mathematics, 95*, 195–205.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates. *Journal of Experimental Psychology: General, 141*(1), 2–18.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). *Multivariate data analysis*. Harlow: Pearson.

Hodson, D. (2014). Learning science, learning about science, doing science. *International Journal of Science Education, 36*, 2534–3255.

Holland, J. L. (1963). Explorations of a theory of vocational choice and achievement. *Psychological Reports, 12*, 547–594.
Hubley, A. M., & Zumbo, B. D. (2017). Response processes in the context of validity. In B. D. Zumbo & A. M. Hubley (Eds.), *Understanding and investigating response processes in validation research* (pp. 1–12). Springer.
Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika, 39*, 31–36.
Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.
KMK (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der BRD). (Ed.). (2020). *Bildungsstandards im Fach Biologie für die Allgemeine Hochschulreife (Biology education standards for the Allgemeine Hochschulreife)*. Luchterhand.
KMK (Ed.). (2005). *Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss (Biology education standards for the Mittlere Schulabschluss)*. Munich: Wolters Kluwer.
KMK (Ed.). (2019). *Ländergemeinsame inhaltliche Anforderungen für die Fachwissenschaften und Fachdidaktiken in der Lehrerbildung (Common national content requirements for the subject sciences and educational subjects in teacher training)*. Received from https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2008/2008_10_16-Fachprofile-Lehrerbildung.pdf. Accessed 21 Jan 2022.
Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T., & Hachfeld, A. (2013). Professional competence of teachers. *Journal of Educational Psychology, 105*, 805–820.
Lamminpää, J., Vesterinen, V. M., & Puutio, K. (2020). Draw-a-science-comic. *Research in Science & Technological Education*, 1–22. https://doi.org/10.1080/02635143.2020.1839405
Liang, L. L., Chen, S., Chen, X., Kaya, O. N., Adams, A. D., Macklin, M., & Ebenezer, J. (2008). Assessing preservice elementary teachers' views on the nature of scientific knowledge: A dual-response instrument. *Asia-Pacific Forum on Science Learning and Teaching, 9*(1), 1–20.
Losh, S. C., Wilke, R., & Pop, M. (2008). Some methodological issues with "draw a scientist tests" among young children. *International Journal of Science Education, 30*, 773–792.
McComas, W. F., & Clough, M. P. (2020). Nature of science in science instruction. In W. F. McComas (Ed.), *Nature of Science in Science Instruction Rationales and Strategies* (pp. 3–22). Springer.
Miele, E. (2014). Using the draw-a-scientist test for inquiry and evaluation. *Journal of College Science Teaching, 43*(4), 36–40.
Milford, T. M., & Tippett, C. D. (2013). Preservice teachers' images of scientists. *Journal of Science Teacher Education, 24*, 745–762.
Miller, D. I., Nolla, K. M., Eagly, A. H., & Uttal, D. H. (2018). The development of children's gender-science stereotypes. *Child Development, 89*, 1943–1955.
NGSS Lead States. (2013). *Next generation science standards*. National Academies.
NSTA & ASTE (2020). *Standards for science teacher preparation*. Received from https://static.nsta.org/pdfs/2020NSTAStandards.pdf. Accessed 21 Jan 2022.
Reinisch, B., Krell,. M., Hergert, S., Gogolin, S., & Krüger, D. (2017). Methodical challenges concerning the Draw-A-Scientist Test: a critical view about the assessment and evaluation of learners' conceptions of scientists. *International Journal of Science Education*, *39*, 1952–197.
Roberts, D. A. (2007). Scientific literacy/science literacy. In N. G. Lederman & S. K. Abell (Eds.), *Handbook of research on science education* (pp. 729–778). Erlbaum.
Treagust, D. F., & Duit, R. (2008). Conceptual change. *Cultural Studies of Science Education, 3*, 297–328.
Urhahne, D., Kremer, K. & Mayer, J. (2011). Conceptions of the nature of science – are they general or context specific? *International Journal of Science and Mathematics Education , 9*, 707–730.
van Tuijl, C., & van der Molen, J. H. W. (2016). Study choice and career development in STEM fields. *International Journal of Technology and Design Education*, *26*, 159–183.
Wentorf, W., Höffler, T. N., & Parchmann, I. (2015). Schülerkonzepte über das Tätigkeitsspektrum von Naturwissenschaftlerinnen und Naturwissenschaftlern (Students' concepts about scientists' activities). *Zeitschrift für Didaktik der Naturwissenschaften, 21*, 207–222.