PRZEGLĄD PSYCHOLOGICZNY – THE REVIEW OF PSYCHOLOGY
2024, TOM 67, NR 3, 115–128
DOI: 10.31648/przegldpsychologiczny.10476

Bremer Key: The Development and Validation of a Child-Sensitive Language Competence Test for Teachers

Tamas Rotschild[1]
The University of Bremen, Germany
Faculty 12: Pedagogy and Educational Sciences
https.//orcid.org/0000-0001-9244-2815

Abstract

Aim: Language is a powerful instrument of educators, yet it is anything but neutral, carrying nuanced meanings that extend beyond the literal. These nuances, laden with emotions and evaluations, exert a cumulative impact on the inner worlds of children. Therefore, a high level of child-sensitive language (CSL) competence is essential for teachers.

Method: To enhance prospective teachers' language competence, a Situational Judgement Test (SJT) was developed and validated as part of doctoral research at the University of Bremen, Germany. The test items and response options were developed by a team of three in-service teachers, led by the doctoral researcher, and supported by two university professors with expertise in the field, following established development practices.

Results: Pilot testing involved 47 in-service teachers from multiple public schools and 55 university students across three institutions. Following iterative refinement, the test was finalized with 13 items, demonstrating strong internal consistency indicated by a computed value of $\omega = .81$. The test and its utility were positively evaluated by the participants.

Conclusion: The study concludes that this instrument can be utilized across various educational domains and teacher training programs to enhance educational practices and outcomes.

Keywords: language competence, child's perspective, Situational Judgment Test, educational assessment, teacher training

[1] Correspondence addresses: tam_rot@uni-bremen.de; rotschild.tommy@gmail.com.

The *Bremer Key* is an assessment tool designed to evaluate pre-service teachers' competence in applying specific semantic and pragmatic choices that shape the linguistic and psychological properties of child-sensitive communication, as outlined by Rotschild (2024). According to Rotschild, messages constructed and communicated in CSL have a positive impact on children's intrapsychic world, supporting their emotional well-being, sense of self, and cognitive development. Teachers' regular comments and feedback on children's attributes, performance, and behavior throughout a typical school day exert considerable power over children's development (Johnston, 2004). Their messages are primary sources of children's psychological experiences that help children make sense of themselves and form their self-view (Burns, 1982). The significance of holding positive views of oneself cannot be overstated. Research findings indicate that positive self-concept is crucial in the development and preservation of mental health and has a far-reaching effect on various important psychological, behavioral, and educational results (Craven & Marsh, 2008; Ybrandt, 2007). While language is a powerful instrument for shaping self-perceptions and behavior without force, it is anything but neutral. Every spoken word and phrase convey meanings beyond the literal, carrying nuances of emotions and evaluations. They do not have an immediate magical effect, but rather a latent cumulative impact, both individually and in the way sentences are constructed from them (Bolinger, 1980). Teachers' messages unwittingly inform children about what kinds of people the teacher think they are and encourage the children to practice being those kinds of people (Johnston, 2004). Today's messages become a child's self-concept tomorrow. It is therefore of utmost importance for teachers to consciously choose language that empowers children and allows them to define themselves favorably. This paper discusses the conceptual development and validation of the *Bremer Key* – an SJT adapted to assess teachers' CSL competence and predict their communication behavior when addressing children in sensitive and challenging situations. The administration of the *Bremer Key* can provide data to help develop instructional materials and training courses aimed at improving prospective teachers' ability to foster a positive, supportive learning and developmental environment for children through effective language use. In addition to its potential utility in teacher training, the *Bremer Key* holds promise for broader applications. It could be utilized in ongoing professional development for in-service teachers, contributing to continuous improvement in classroom communication practices. Moreover, its application in curriculum design for teacher training programs could help ensure that aspiring educators are better prepared to engage effectively with students. For teacher recruitment and selection, educational institutions could integrate this instrument into the hiring process. Through a systematic assessment of candidates' language competence and communication behavior tendencies, it could assist in selecting individuals who could potentially demonstrate the essential skills for effective educational practices. Furthermore, the instrument could be instrumental in policy development within the education sector. Policymakers could use the results to potentially shape informed policies aimed at improving overall teaching quality, with a specific emphasis on fostering effective language

use and communication in the classroom. In essence, this instrument has the potential to be a key factor in ongoing efforts to enhance educational practices and outcomes.

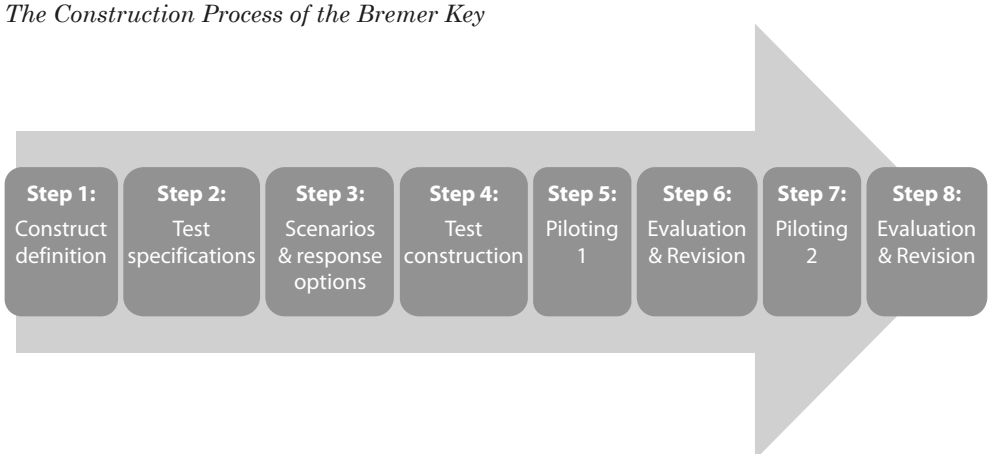
Method

Developing a Situational Judgement Test to Assess Teachers' Language Competence

Adopting established development practices (Delgado-Rico et al., 2012; Patterson et al., 2015; Pollard & Cooper-Thomas 2015; Whetzel et al., 2020) and drawing upon contemporary research findings (Reiser et al., 2022; Smith et al., 2020, 2022), the construction of the *Bremer Key* proceeded in eight steps, as shown in Figure 1 below. Participants included experts, in-service teachers, and German university students from faculties of education or affiliated disciplines (e.g., psychology, school social work). Throughout all phases, participants were informed about the purpose of the study. Participation was voluntary. The tests were anonymized; examinees were not required to provide any personal identifiers, ensuring their privacy. Their IP addresses were neither recorded nor stored.

Figure 1

The Construction Process of the Bremer Key



Step 1: Construct Definition

CSL, the construct to be assessed, was conceptualized based on the research findings presented in Rotschild (2023), as illustrated in Table 1 below (p. 118) adopted from Delgado-Rico et al. (2012).

Table 1

Conceptualization of the Construct "Child-Sensitive Language" and its Key Components

Child-Sensitive Language (CSL)
The language crafted to consider children's cognitive abilities, individuality, emotional sensitivity, and self-perception formation. It respects children's unique experiences and perspectives, aiming to nurture their self-esteem, confidence, and positive self-image.
CSL_1 Language that demonstrates an understanding of and sensitivity to a child's emotions and experiences.
CLS_2 Language that refrains from making value judgments or criticisms.
CLS_3 Language that avoids using words, phrases, or tones that may induce fear or anxiety in children.
CLS_4 Language that expresses genuine appreciation and encouragement towards children.
CLS_5 Language that offers words of support, reassurance, and guidance to children.

CSL is a higher-level, latent, multidimensional construct as defined by Law et al. (1998). It integrates several related but distinct lower-level, stand-alone constructs, each with its own set of observable indicators (Johnson et al., 2012). Each of these constructs contributes uniquely to the overarching concept of CSL, making it richer and more meaningful than the individual dimensions considered separately. The interaction between the different dimensions can create synergies that enhance the overall functionality and impact of the construct (Carver, 1989).

Step 2: Test Specifications

In this phase, a comprehensive framework was established for the development of the assessment. The specifications crafted encompassed the target audience, purpose of testing, and chosen test method. Given their pivotal role in shaping children's self-perceptions through their comments on academic and behavioral performance at school, teachers, psychologists, and school social workers were identified as the target audience. Recognizing the centrality of language in their profession, the purpose of the test was meticulously defined as assessing these practitioners' CSL competence. To operationalize this purpose effectively, the SJT with multiple-choice items was selected as the test method due to its established efficacy in providing a valid, reliable, and evidence-based assessment of non-academic attributes, including communication, essential for effective educational work (Al Hashmi & Klassen, 2020; Durksen & Klassen, 2018). SJTs present examinees with work-related scenarios and a list of potential actions to choose from. Test takers are typically required to select the best possible action or the one they would most likely perform in a given situation (McDaniel & Nguyen, 2001; McDaniel et al., 2007). SJTs offer several strengths, including their ease of administration to large groups of examinees simultaneously over the internet, strong psychometric properties, and favorable perceptions among test takers (Lievens et al., 2008).

Step 3: Scenarios and Response Options

At this stage, 36 real-life scenarios, each with 3 options, were drafted by three in-service teachers and reviewed by two experts to ensure practical relevance and accuracy. The distractors, re-formulated for test usage, were drawn from real-world practices (praxis) to reflect common misconceptions or misunderstandings encountered in professional settings. This approach maintains the test's authenticity and relevance to the real-life challenges practitioners face. The keys, or correct answers, were informed by Rotschild's (2024) child-sensitive communication guidelines, grounding the correct responses in empirical evidence. As a result, the scenarios and options directly assess competence in applying child-sensitive communication strategies, making the test a robust tool for evaluating professionals in this domain. The writing of the test items followed the suggestions outlined by Haladyna et al. (2002) The decision to use three choices per item was made in light of empirical findings (Rodriguez, 2005). Table 2 below shows a typical item.

Table 2

Sample Test Item

In class, the students have been given an assignment to work on a set of math exercises. Leon, however, is sitting at his desk and hasn't started his work. He seems disengaged, looking around the classroom while other students are working.	
a	Leon, what can I do to help you get started on your exercises?
b	Leon, why aren't you doing the exercises?
c	Leon, would you like to start doing the exercises?

Step 4: Test Construction

The test items were assessed for representativeness and content similarity, leading to the exclusion of 12 items. The remaining 24 items were then split into two sets, Version A and Version B. Each item in Version A had a matching item in Version B, resulting in 24 unique questions distributed across the two versions. The keys were distributed among different options, ensuring an equal number of occurrences in each option position, as advocated by Haladyna and Downing (1985). The distribution of correct answer positions was balanced and randomized to avoid answer position clustering.

Step 5: Piloting 1

Each version was administered in a face-to-face manner in April 2023, using a paper-and-pencil format to a group of 20 in-service teachers ("expert examinees") attending a professional development program at university. Of the 20 participants, 11 were female and 9 were male. Their experience ranged from 4 to 18 years (M = 7.8).

Step 6: Evaluation and Revision

Evaluation Phase

The aim in this phase was twofold: (1) selecting test items from the two versions for inclusion in a single test and (2) revising and modifying them for optimal efficacy. The evaluation commenced with item analysis to assess item difficulty (*p*). Although the statistical power is limited with a sample size of 10, calculating the difficulty level for each item provided valuable preliminary insights into item performance, helping to identify potential issues for revision. The results are shown in Table 3.

Table 3

Item Analysis Data

Item (I)	Version A (*p*)	Version B (*p*)
1	.30	.30
2	.20	.70
3	.70	.00
4	.30	.20
5	.60	.20
6	.10	.40
7	.30	.30
8	.00	.50
9	.20	.60
10	.10	.40
11	.60	.10
12	.00	.40

The analysis data was enhanced by comments from a university professor with expert knowledge in the field and by 13 of the 20 examinees who participated in a voluntary open group discussion held a week after the test administration. As a result, eight items (1, 2, 3, 4, 5, 7, 8, and 12) from Version A and five items (1, 2, 4, 5, and 12) from Version B were selected as prospective items for inclusion in the second draft. The selection criteria included clarity, low cognitive complexity, and typicality of the scenarios described by the items. A 14th item arose from the examinees' comments, leading to its development and eventual inclusion in the revised draft. The decision to limit the number of items to 14 on the test was made to make it more convenient for participants, reduce fatigue, facilitate easy administration, scoring, and interpretation.

Revision Phase

The items were revised based on the comments and insights gained from the review of further sources (Boland et al., 2010; Holland & Stevens 2021; Mayer

& Moreno, 2003; Zhang et al., 2022) to improve the balance in the difficulty distribution, ensure an even distribution of discriminative items and valid assessment. Names were gender-neutralized where linguistic context allowed to avoid biases associated with specific genders. Stereotypes were mitigated to ensure ethicality and promote accuracy in assessment. Language was simplified to minimize ambiguity. To improve clarity and focus, item length was shortened and extraneous information was removed from stems as well as from response options. Additionally, distractors were adjusted or replaced.

A single revised test was constructed with LimeSurvey and stored on server of the University of Bremen. To the 14 domain specific items, 5 feedback questions were added to collect evaluative feedback from the participants:

1. Did you find the scenarios presented in the test realistic and representative of actual communication situations in the classroom?
2. Were the response options provided in the test relevant and appropriate for the given scenarios?
3. Did you find the response options plausible in the given scenarios?
4. Were the test items clearly written?
5. Do you think the test can assess whether the test-taker can differentiate between positively and negatively formulated messages?

Each question offered a five-point-scale answer option with numerical value: (a) *Yes*/ *Very well* [5], (b) *Mostly*/ *Well* [4], (c) *Moderately* [3], (d) *Rather not*/ *Poorly* [2], and (e) *Not at all* [1]. The test concluded with an open-ended question: "How did you find the content and structure of this test? Did it prompt you to reflect on your own communication? What did you learn from the test?" to request further input from the test-takers.

Step 7: Piloting 2

The second piloting took place online in November – December 2023. The settings were configured to prevent test-takers from skipping items, logging out and in again, or taking the test twice. No time limit for completing the test was set.

The test-takers ($N = 82$) comprised 27 in-service teachers and 55 university students studying education or related disciplines (e.g., psychology) at three German institutions. The in-service teachers were recruited via direct email, while the students received the test link indirect from their professors.

Step 8: Evaluation and Revision

Evaluation Phase

The evaluation at this stage served to validate the changes, helping to confirm that they effectively enhance the reliability and clarity of the items. Additionally, the analysis aimed to identify any residual or new issues that may have arisen, ensuring that the items are error-free and build a reliable and valid assessment instrument.

Results

Item Analysis

The items were analyzed to determine their level of difficulty (*p*) and discriminative power (*D*). The results are shown in Table 4.

Table 4

2nd Item Performance Analysis Data

Item	Original Item	*p*	*D*
1	A1	.30	.50
2	B5	.50	.20
3	A5	.70	.30
4	A2	.00	–.10
5	B2	.70	.10
6	A4	.70	.70
7	A8	.80	.50
8	B1	.20	.40
9	B4	.70	.50
10	A3	.50	.20
11	–	.50	.40
12	A12	.50	.60
13	B12	.60	.50
14	A7	.60	.00

Note. A and B refer to the test version in Piloting 1.

All items except for Items 4 and 7 fall within a midrange of difficulty ($.30 < p < .70$) and therefore have the potential to provide differential information. With most items exceeding the accepted value (.30) indicating reasonably good discriminative power, the diverse discrimination indices of the items enable a thorough assessment across a wide spectrum of abilities (Ebel & Frisbie, 1991). Alongside their midrange difficulty levels, they contribute to a well-balanced test construction. That is, an item with moderate difficulty and lower discrimination can complement other items that are more challenging and have higher discrimination. However, it's noteworthy that Item 14 has a discrimination index of .00. The negative discrimination index of Item 4 likely stems from social desirability bias. This bias occurs when participants are more inclined to choose response options that align with societal norms or present themselves in a favorable light, rather than selecting the correct answer, which may be less socially desirable. The combination of high incorrect response rates and the

resulting difficulty level further reinforces the negative discrimination index observed for Item 4. The low discrimination index observed for Items 2, 5, and 10 can be attributed to their focus on assessing core knowledge. While these items may not effectively differentiate between high and low performers, they play a crucial role in evaluating whether test-takers possess fundamental understanding required for the subject. Additionally, the low discriminative power of Items 2 and 10 may be exacerbated by their nonfunctional distractors.

Distractor Analysis

Based on Berk (1984) and Raymond et al. (2019), seven distractors were regarded as nonfunctional for being selected by fewer than 5% of the examinees.

Reliability

Coefficient omega (ω) was computed to estimate reliability. Unlike the widely used coefficient alpha (α), which assumes tau-equivalence (τ) (equal factor loadings), coefficient omega accounts for varying degrees of item-factor relationships inherent in multidimensional constructs by allowing for different factor loadings and item-specific errors, thus providing a more accurate and confident estimate of the reliability of scales where items may vary in their contribution to the underlying factors (Gignac, 2013; Hayes & Coutts, 2020; Watkins, 2017). The computed value of $\omega = .81$ exceeds the comparable benchmark of $\alpha = .7$ for good internal consistency (Kline, 2000).

The computation involved three distinct steps, utilizing the functionalities of SPSS and Microsoft Excel in tandem:

1. A factor analysis using the maximum likelihood method was performed, extracting a single factor to assess the correlation of each test item with the construct under scrutiny. Item 14 was excluded from further analysis due to its low factor loading (.025).
2. A factor analysis using the maximum likelihood method was performed to extract five factors, corresponding to the different facets of the construct, and to identify the primary (highest) loading of each item.
3. Coefficient ω was computed using the following formula:

$$\omega = \frac{(\Sigma\lambda)^2}{(\Sigma\lambda)^2 + (\Sigma\varepsilon)}$$

where $(\Sigma\lambda)^2$ is the summed primary factor loadings squared and $(\Sigma\varepsilon)$ is the summed error variance.

By computing reliability for the overarching construct using each item's highest loading, the reliability assessment focused on the primary construct of interest. Including only the highest loadings maximized each item's contribution to reliability and enhanced construct validity, as these items have a stronger association with the underlying construct. High-loading items also ensure that the construct is well-represented across its various dimensions, thereby supporting

content validity, too. Nevertheless, this approach can lead to the loss of multifactor contributions. Items that load on multiple factors may only contribute to the composite score based on their highest loading, potentially obscuring their relevance to secondary factors. Finally, this method might overlook the broader contributions of items with moderate but meaningful loadings.

Validity

The validity of the instrument is ensured through multiple avenues. Firstly, the instrument's design followed established procedures. The construct under assessment is clearly conceptualized and grounded in empirical evidence. Scenarios, reflective of real-life classroom situations and drawn from firsthand experiences of practitioners ("insiders"), imbue the instrument with practical applicability. The response options were meticulously crafted based on the operationalization of the construct, as presented in Rotschild. In addition, the test items underwent review, evaluation, and approval by two experts. Furthermore, the validation process was enriched by invaluable insights gleaned from the first piloting phase, where examinees' feedback offered qualitative perspectives on item properties. The second piloting phase of the instrument yielded crucial statistical data, providing key indicators supporting the test's overall validity. The test-takers (N = 82) rated the items as clearly written, with a mean score of 4.86 (SD = 0.56), indicating a high level of clarity and supporting the test's content and face validity. Additionally, the scenarios described by the test items were evaluated as authentic, with a mean of 3.97 (SD = 0.82), further corroborating the test's content validity. The response options were deemed plausible, scoring a mean of 4.13 (SD = 0.83), which is essential for ensuring that the test measures realistic decision-making situations, reinforcing content validity. The relevance and appropriateness of the scenarios and response options were also affirmed with a mean score of 3.76 (SD = 0.79), contributing to both content and construct validity. Finally, the test was perceived as effective in assessing child-sensitive communication competence, achieving a mean score of 3.94 (SD = 0.73), which further contributes to construct validity.

Utility

Among the various subjects available for discussion, a significant portion of examinees – 35 individuals, constituting 43% – opted to report, in response to the open-ended question at the end of the test, that they were prompted to reflect on their own language usage and/or think over the significance of linguistic subtleties. This acknowledgment underscores the test's practicality and potentiality as a tool to enhance communication competence.

Revision Phase

Based on its psychometric data, complexity compared to other items, and similarity in content with them, Item 14 was excluded. The seven nonfunctional

distractors were replaced. Item 4 was retained unchanged, along with Items 2, 5, and 10, for their diagnostic potential and coverage of core skills within the domain being tested. The 13 items constitute the final version of the instrument to be used in large-scale assessments.

Discussion

The development and validation of the *Bremer Key* represent a significant advancement in enhancing teacher training programs. The SJT demonstrates strong content validity and reliability for assessing both teacher trainees' and in-service teachers' CSL competence. The balance in item difficulty and discriminative power ensures that the test is both challenging and fair, providing a nuanced assessment of examinees' abilities to effectively navigate child-sensitive interactions. Feedback from examinees confirmed the authenticity of the test scenarios and the plausibility of the response options. Additionally, the test's utility in raising awareness of the importance of linguistic nuances and in encouraging reflection on one's own language usage has been reinforced. With its robust psychometric properties and applicability, this tool holds significant promise for improving teaching practices and ultimately advancing educational outcomes. The *Bremer Key* is meant to be administered online to teacher trainees across various universities, focusing on five key domains of CSL use. Trainees' performance is expected to be assessed against a predetermined cut-off score, reflecting a mastery level justifiably expected from effective practitioners. Once the data are collected, descriptive statistics will be applied to summarize overall performance, including mean scores, standard deviation, and the proportion of trainees who meet the required competence level. The results will then be compiled into detailed reports and shared with heads of faculty at each university. These reports are intended to provide insight into areas where trainees excel and where further development is needed. Faculty members will be encouraged to use this data to gain information on the development of profession-specific communication education, allowing for the creation of targeted programs that enhance the professional qualities of the trainees. Indeed, a significant limitation of the *Bremer Key* is its inability to capture non-verbal aspects of communication, such as body language, facial expressions, and gestures. Consequently, the test's predictive validity in real-world teaching scenarios may be compromised. By focusing primarily on sentence-level responses to isolated scenarios, the *Bremer Key* may not fully capture teachers' discourse-level communication skills, including their ability to utilize CSL in ongoing interactions.

In essence, while the *Bremer Key* assesses certain aspects, it falls short of comprehensively addressing the complex nature of the broader communication skills required in educational contexts. However, it can serve as an effective diagnostic tool to provide preliminary results and can contribute significantly to enriching teacher training programs. To enhance the predictive validity of the *Bremer Key*, future adaptations or supplementary assessments could consider

incorporating methods that evaluate both verbal and non-verbal communication cues. Furthermore, future research should focus on exploring the implementation and effectiveness of the *Bremer Key* in broader educational settings to further enhance its utility and impact.

References

Al Hashmi, W. A., & Klassen, R. M. (2020). Developing a situational judgement test for ad mission into initial teacher education in Oman: An exploratory study. *International Journal of School & Educational Psychology*, *8*(Supplement 1), 187–198. https://doi.org/10.1080/21683603.2019.1630042

Berk, R. A. (1984). Conducting the Item Analysis. In R. A. Berk (Ed.), *A Guide to criterion-referenced test construction* (pp. 97–143). The John Hopkins University Press.

Boland, R. J., Lester, N. A., & Williams, E. (2010). Writing multiple-choice questions. *Academic Psychiatry*, *34*, 310–316. https://doi.org/10.1176/appi.ap.34.4.310

Bolinger, D. (1980). *Language – The loaded weapon*. Longman.

Burns, R. B. (1982). *Self-concept development and education*. Holt.

Carver, C. S. (1989). How should multifaceted personality constructs be tested? Issues illustrated by self-monitoring, attributional style, and hardiness. *Journal of Personality and Social Psychology*, *56*(4), 577–585. https://doi.org/10.1037/0022-3514.56.4.577

Craven, R. G., & Marsh, H. W. (2008). The centrality of the self-concept construct for psychological well-being and unlocking human potential: Implications for child and educational psychologists. *Educational & Child Psychology*, *25*(2), 104–118. http://dx.doi.org/10.53841/bpsecp.2008.25.2.104

Delgado-Rico, E., Carrctero-Dios, H., & Ruch, W. (2012). Content validity evidences in test development: An applied perspective. *International Journal of Clinical and Health Psychology*, *12*(3), 449–460. https://doaj.org/article/e3c64d10e477402ca1f9add91cb467d5

Durksen, T. L., & Klassen, R. M. (2018). The development of a situational judgement test of personal attributes for quality teaching in rural and remote Australia. *Australian Educational Researcher*, *45*(2), 255–276. https://doi.org/10.1007/s13384-017-0248-5

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Allyn & Bacon.

Gignac, G. E. (2013). On the inappropriateness of using items to calculate total scale score reliability via coefficient alpha for multidimensional scales. *European Journal of Psychological Assessment*, *30*(2), 130–139. https://doi.org/10.1027/1015-5759/a000181

Haladyna, T. M., & Downing, S. M. (5–9 April, 1988). *Functional distractors: Implications for test-item writing and test design* [paper presentation]. Annual Meeting of the American Educational Research Association, New Orlean, USA. http://files.eric.ed.gov/fulltext/ED293851.pdf

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. Applied Measurement in Education, *15*(3), 309–334. https://doi.org/10.1207/S15324818AME1503_5

Haladyna, T. M., Downing, S. M., & Steven, M. (31 March–4 April, 1985). *A quantitative review of research on multiple-choice item writing* [paper presentation]. Annual Meeting of the American Educational Research Association, Chicago, USA. https://eric.ed.gov/?id=ED255580

Hayes, A. F., & Coutts, J. J. (2020). Use Omega Rather than Cronbach's Alpha for Estimating Reliability. But…. *Communication Methods and Measures*, *14*(1), 1–24. https://doi.org/10.1080/19312458.2020.1718629

Holland, J., & Stevens, N. (2021). *Guidelines for the development of multiple-choice items & assessments*. RCSI University of Medicine and Health Sciences.

Johnston, P. (2004). *Choice words*. Stenhouse.

Johnson, R. E., Rosen, C. C., Chang, Ch., H., Djurdjevic, E., & Taing, M. U. (2012). Recommendations for improving the construct clarity of higher-order multidimensional constructs. *Human Resource Management Review*, *22*(2), 62–72. https://doi.org/10.1016/j.hrmr.2011.11.006

Kline, P. (2000). *The handbook of psychological testing* (2nd ed.). Routledge.

Law, K. S., Wong, C., & Mobley, W. M. (1998). Toward a taxonomy of multidimensional constructs. *The Academy of Management Review*, *23*(4), 741–755. https://doi.org/10.5465/amr.1998.1255636

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review*, *37*(4), 426–441. https://doi.org/10.1108/00483480810877598

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, *38*(1), 43–52. https://doi.org/10.1207/s15326985ep3801_6

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. (2007). Situational judgement tests, response instructions, and validity: A meta-analysis. *Personnel Psychology*, *60*(1), 63–91. https://doi.org/10.1111/j.1744-6570.2007.00065.x

McDaniel, M. A., & Nguyen, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment*, *9*(1–2), 103–113. https://doi.org/10.1111/1468-2389.00

Patterson, F., Zibarras, L., & Ashworth, V. (2015). Situational judgement tests in medical education and training: Research, theory and practice: AMEE Guide No. 100. *Medical Teacher*, *38*(1), 3–17. https://doi.org/10.3109/0142159x.2015.1072619

Pollard, S., & Cooper-Thomas, H. D. (2015). Best practice recommendations for Situational judgment tests. *Australasian Journal of Organisational Psychology*, *8*, e7. https://doi.org/10.1017/orp.2015.6

Raymond, M. R., Stevens, C., & Bucak, S. D. (2019). The optimal number of options for multiple-choice questions on high-stakes tests: Application of a revised index for detecting nonfunctional distractors. *Advances in Health Sciences Education*, *24*(1), 141–150. https://doi.org/10.1007/s10459-018-9855-9

Reiser, S., Schacht, L., Thomm, E., Figalist, C., Janssen, L., Schick, K., Dörfler, E., Berberat, P. O., Gartmeier, M., & Bauer, J. (2022). A video-based situational judgement

test of medical students' communication competence in patient encounters: Development and first evaluation. *Patient Education and Counseling*, *105*(5), 1283–1289. https://doi.org/10.1016/j.pec.2021.08.020

Rodriguez, M. C. (2005). Three options are optimal for Multiple-Choice items: a Meta-Analysis of 80 years of research. *Educational Measurement*, *24*(2), 3–13. https://doi.org/10.1111/j.1745-3992.2005.00006.x

Rotschild, T. (2023). Why and how to foster learning-disabled children's emotional intelligence? *Insights into Learning Disabilities*, *20*(2), 153–175. https://files.eric.ed.gov/fulltext/EJ1401946.pdf

Rotschild, T. (2024). The impact of communication: A practical guide for teachers in fostering positive self-concept in children with learning disabilities. *Journal of Research in Special Educational Needs*, *00*, 1–13. https://doi.org/10.1111/1471-3802.12709

Smith, K. J., Flaxman, C., Farland, M. Z., Thomas, A., Buring, S. M., Whalen, K., & Patterson, F. (2020). Development and validation of a situational judgement test to assess professionalism. *American Journal of Pharmaceutical Education*, *84*(7), Article 7771, 985–992. https://doi.org/10.5688/ajpe7771

Smith, K. J., Neely, S., Dennis, V. C., Miller, M. M., & Medina, M. S. (2022). Use of situational judgment tests to teach empathy, assertiveness, communication, and ethics. *American Journal of Pharmaceutical Education*, *86*(6), Article 8761. https://doi.org/10.5688/ajpe8761

Watkins, M. W. (2017). The reliability of multidimensional neuropsychological measures: from alpha to omega. *The Clinical Neuropsychologist*, *31*(6–7), 1113–1126. https://doi.org/10.1080/13854046.2017.1317364

Whetzel, D. L., Sullivan, T. S., & McCloy, R. A. (2020). Situational judgment tests: An overview of development practices and psychometric characteristics. *Personnel Assessment and Decisions*, *6*(1), Article 1, 1–16. https://doi.org/10.25035/pad.2020.01.001

Ybrandt, H. (2007). The relation between self-concept and social functioning in adolescence. *Journal of Adolescence*, *31*(1), 1–16. https://doi.org/10.1016/j.adolescence.2007.03.004

Zhang, N., He, G., Shi, D., Zhao, Z., & Li, J. (2022). Does a gender-neutral name associate with the research impact of a scientist? *Journal of Informetrics*, *16*(1), Article 101251. https://doi.org/10.1016/j.joi.2022.101251