**MULTIPLE-CHOICE TEST ITEMS: SELECTED REFERENCES**

**(Last updated 4 October 2024)**

Ackerman, T. A., & Smith, P. L. (1988). A comparison of the information provided by essay, multiple-choice, and free-response writing tests. *Applied Psychological Measurement*, *12*(2), 117-128.

Adeosun, S. O. (2023). Differences in multiple-choice questions of opposite stem orientations based on a novel item quality measure. *American Journal of Pharmaceutical Education, 87*(2) ajpe8934. doi: 10.5688/ajpe8934.

Albanese, M. A., Kent, T. H., & Whitney, D. R. (1979). Cluing in multiple-choice test items with combinations of correct responses. *Academic Medicine*, *54*(12), 948-50.

Al Fraidan, A., & Al-Khalaf, K. (2012). Test-taking strategies of Arab EFL learners on multiple choice tests. *International Education Studies*, *5*(4), 80-85.

Al-Hamly, M., & Coombe, C. (2005). To change or not to change: Investigating the value of MCQ answer changing for Gulf Arab students. *Language Testing, 22*(4), 509-531. [http://ltj.sagepub.com/content/22/4/509.full.pdf+html](https://mail.middlebury.edu/owa/redir.aspx?C=gr0BLvrcFki4pdfj_1qh6sVKqt1UztEIp1unuG4vGM2L4seXpMkli5TyCnAgO13_Xl0dgDHSwW8.&URL=http%3a%2f%2fltj.sagepub.com%2fcontent%2f22%2f4%2f509.full.pdf%2bhtml)

Ali, S. H., & Ruit, K. G. (2015). The Impact of item flaws, testing at low cognitive level, and low distractor functioning on multiple-choice question quality. *Perspectives on Medical Education*, *4*(5), 244-251.

Amini, M., & Ibrahim-González, N. (2012). The washback effect of cloze and multiple choice test items on vocabulary acquisition. *Language in India*, *12*(7), 71-91.

Arabmofrad, A., & Mehdiabadi, F. (2022). Developing a multiple-choice discourse completion test for Iranian EFL learners: The case of the four speech acts of apology, request, refusal and thanks. *Language Related Research*, *13*(4), 1-26.

Attali, Y., & Bar‐Hillel, M. (2003). Guess where: The position of correct answers in multiple‐choice test items as a psychometric variable. *Journal of Educational Measurement*, *40*(2), 109-128.

Ayanwale, M. A., & Ndlovu, M. (2021). Ensuring scalability of a cognitive multiple-choice test through the Mokken package in R programming language. *Education sciences*, *11*(12), 794.

Bacon, D. R. (2003). Assessing learning outcomes: A comparison of multiple-choice and short-answer questions in a marketing context. *Journal of Marketing Education*, *25*(1), 31-36.

Bailey, K. M., & Curtis, A. (2015). *Learning about language assessment: Dilemmas, decisions and directions* (2nd ed.). National Geographic Learning.

Becker, W. E., & Johnston, C. (1999). The relationship between multiple choice and essay response questions in assessing economics understanding. *Economic Record*, *75*(4), 348-357.

Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free‐response and multiple‐choice items. *Journal of Educational Measurement*, *28*(1), 77-92.

Ben‐Shakhar, G., & Sinai, Y. (1991). Gender differences in multiple‐choice tests: the role of differential guessing tendencies. *Journal of Educational Measurement*, *28*(1), 23-35.

Berg, C. A., & Smith, P. (1994). Assessing students' abilities to construct and interpret line graphs: Disparities between multiple‐choice and free‐response instruments. *Science Education*, *78*(6), 527-554.

Betts, L. R., Elder, T. J., Hartley, J., & Trueman, M. (2009). Does correction for guessing reduce students’ performance on multiple‐choice examinations? Yes? No? Sometimes?. *Assessment & Evaluation in Higher Education*, *34*(1), 1-15.

Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats—it does make a difference for diagnostic purposes. *Applied Psychological Measurement*, *11*(4), 385-395.

Biria, R. E. Z. A., & Dehghan, M. A. H. D. I. (2016). The relationship between self-regulated learning and test method format: The case of multiple choice and open-ended questions in L2 reading. *Journal of Global Research in Education and Social Science*, *7*(3), 160-167.

Bormuth, J. R. (1967). Comparable cloze and multiple-choice comprehension test scores. *Journal of Reading*, *10*(5), 291-299.

Brady, A. M. (2005). Assessment of learning with multiple-choice questions. *Nurse Education in Practice*, *5*(4), 238-242.

Brame, C. J. (2014). *Writing good multiple choice test questions*. Vanderbilt University. http://cft.vanderbilt.edu/guides-sub-pages/writing-good-multiple-choice-test-questions/

Bridgeman, B. (1992). A comparison of quantitative questions in open‐ended and multiple‐choice formats. *Journal of Educational Measurement*, *29*(3), 253-271.

Bridgeman, B., & Lewis, C. (1994). The relationship of essay and multiple‐choice scores with grades in college courses. *Journal of Educational Measurement*, *31*(1), 37-50.

Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment*, *11*(1), 33-63.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment.* McGraw Hill.

Bruno, J. E., & Dirkzwager, A. (1995). Determining the optimal number of alternatives to a multiple-choice test item: An information theoretic perspective. *Educational and Psychological Measurement*, *55*(6), 959-966.

Buck, G., Tatsuoka, K., & Kostin, I. (1997). The subskills of reading: Rule‐space analysis of a multiple‐choice test of second language reading comprehension. *Language Learning*, *47*(3), 423-466.

Burton, R. F. (2004). Multiple choice and true/false tests: reliability measures and some implications of negative marking. *Assessment & Evaluation in Higher Education*, *29*(5), 585-595.

Burton, R. F. (2005). Multiple‐choice and true/false tests: Myths and misapprehensions. *Assessment & Evaluation in Higher Education*, *30*(1), 65-72.

Burton, S. J., Sudweeks, R. R., Merrill, P. F., & Wood, B. (1991). *How to prepare better multiple-choice test items: Guidelines for university faculty*. Brigham Young University Testing Services.

Bush, M. (2001). A multiple choice test that rewards partial knowledge. *Journal of Further and Higher Education*, *25*(2), 157-163.

Butler, A. C., Karpicke, J. D., & Roediger III, H. L. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied*, *13*(4), 273.

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, *36*(3), 604-616.

Cassels, J. R. T. & Johnstone, A. H. (1984). The effect of language on student performance on multiple choice tests in chemistry. *Journal of Chemical Education, 61,* 613-615.

Celce-Murcia, M., Kooshian, G. B., & Gosak, A. J. (1974). Goal: Good multiple-choice language test items. *English Language Teaching 28*(3), 257-262.

Chang, A. C. S., & Read, J. (2013). Investigating the effects of multiple-choice listening test items in the oral versus written mode on L2 listeners' performance and perceptions. *System*, *41*(3), 575-586.

Chehrazad, M. H., & Ajideh, P. (2015). Effects of different response types on Iranian EFL test takers’ performance. *Iranian Journal of Applied Language Studies*, *5*(2), 29-50.

Cheng H. F. (2004). A comparison of multiple-choice and open ended formats for the assessment of listening proficiency in English. *Foreign Language Annals, 37*(4), 544-555. <http://ltj.sagepub.com/content/27/4/471.full.pdf+html>

Chin, H., & Chew, C. M. (2022). Online cognitive diagnostic assessment with ordered multiple-choice items for word problems involving ‘time’. *Education and Information Technologies*, *27*(6), 7721-7748.

Chiramanee, T., & Currie, M. (2010). The effect of the multiple-choice item format on the measurement of knowledge of language structure. *Language Testing, 27*(4), 471-491.

Choi, I. C. (2008). The impact of EFL testing on EFL education in Korea. *Language Testing*, *25*(1), 39-62. <https://journals.sagepub.com/doi/pdf/10.1177/0265532207083744>

Cizek, G. J., & O'Day, D. M. (1994). Further investigation of nonfunctioning options in multiple-choice test items. *Educational and Psychological Measurement*, *54*(4), 861-872.

Considine, J., Botti, M., & Thomas, S. (2005). Design, format, validity and reliability of multiple choice questions for use in nursing research and education. *Collegian*, *12*(1), 19-24.

Crocker, L., & Schmitt, A. (1987). Improving multiple-choice test performance for examinees with different levels of test anxiety. *The Journal of Experimental Education*, *55*(4), 201-205.

Cross, L. H., & Frary, R. B. (1977). An empirical test of Lord's theoretical results regarding formula scoring of multiple‐choice tests. *Journal of Educational Measurement*, *14*(4), 313-321.

Currie, M., & Chiramanee, T. (2010). The effect of the multiple-choice item format on the measurement of knowledge of language structure. *Language Testing, 27*(4), 471-479. [http://ltj.sagepub.com/content/27/4/471.full.pdf+html](https://mail.middlebury.edu/owa/redir.aspx?C=gr0BLvrcFki4pdfj_1qh6sVKqt1UztEIp1unuG4vGM2L4seXpMkli5TyCnAgO13_Xl0dgDHSwW8.&URL=http%3a%2f%2fltj.sagepub.com%2fcontent%2f27%2f4%2f471.full.pdf%2bhtml)

Daneman, M., & Hannon, B. (2001). Using working memory theory to investigate the construct validity of multiple-choice reading comprehension tests such as the SAT. *Journal of Experimental Psychology: General*, *130*(2), 208.

Danpradit, P., Kongkumnerd, T., & Boonplian, P. (2021). Forwarding strategies to online multiple-choice tests. *Journal of Green Learning*, *1*(2), 52-60.

Davis, F. B. (1959). Estimation and use of scoring weights for each choice in multiple-choice test items. *Educational and Psychological Measurement*, *19*(3), 291-298.

Dehnad, A., Nasser, H., & Hosseini, A. F. (2014). A comparison between three-and four-option multiple choice questions. *Procedia-Social and Behavioral Sciences*, *98*, 398-403.

Delgado, A. R., & Prieto, G. (2003). The effect of item feedback on multiple‐choice test responses. *British Journal of Psychology*, *94*(1), 73-85.

de Winter, J. C. (2023). Can ChatGPT pass high school exams on English language comprehension?. *International Journal of Artificial Intelligence in Education*, 1-16.

Dinçer, B. H., Antonova-Unlu, E., & Kumcu, A. (2022). Assessing the use of multiple-choice translation items in English proficiency tests: The case of the national English proficiency test in Turkey. *Applied Linguistics Review*, *13*(4), 461-475.

Ding, L., & Beichner, R. (2009). Approaches to data analysis of multiple-choice questions. *Physical Review Special Topics-Physics Education Research*, *5*(2), 020103.

Divgi, D. R. (1986). Does the Rasch model really work for multiple choice items? Not if you look closely. *Journal of Educational Measurement*, *23*(4), 283-298.

Dolly, J. P., & Williams, K. S. (1986). Using test-taking strategies to maximize multiple-choice test scores. *Educational and Psychological Measurement*, *46*(3), 619-625.

Downing, S. M. (2002). Construct-irrelevant variance and flawed test questions: Do multiple-choice item-writing principles make any difference?. *Academic Medicine*, *77*(10), S103-S104.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement*, *19*(2), 143-166.

Dressel, P. L., & Schmid, J. (1953). Some modifications of the multiple-choice item. *Educational and Psychological Measurement*, *13*(4), 574-595.

Dudley, A. (2006). Multiple dichotomous-scored items in second language testing: Investigating the multiple true-false item type under norm-referenced conditions. *Language Testing, 23*(2), 198-227. [http://ltj.sagepub.com/content/23/2/198.full.pdf+html](https://mail.middlebury.edu/owa/redir.aspx?C=gr0BLvrcFki4pdfj_1qh6sVKqt1UztEIp1unuG4vGM2L4seXpMkli5TyCnAgO13_Xl0dgDHSwW8.&URL=http%3a%2f%2fltj.sagepub.com%2fcontent%2f23%2f2%2f198.full.pdf%2bhtml)

Ellsworth, R. A., Dunnell, P., & Duell, O. K. (1990). Multiple-choice test items: What are textbook authors telling teachers? *The Journal of Educational Research*, *83*(5), 289-293.

Entin, E. B., & Klare, G. R. (1978). Some inter-relationships of readability, cloze and multiple choice scores on a reading comprehension test. *Journal of Reading Behavior*, *10*(4), 417-436.

Espinosa, M. P., & Gardeazabal, J. (2010). Optimal correction for guessing in multiple-choice tests. *Journal of Mathematical Psychology*, *54*(5), 415-425.

Farley, J. K. (1989). The multiple-choice test: Writing the questions. *Nurse Educator*, *14*(6), 10-12.

Farr, R., Pritchard, R., & Smitten, B. (1990). A description of what happens when an examinee takes a multiple‐choice reading comprehension test. *Journal of Educational Measurement*, *27*(3), 209-226.

Frary, R. B. (1980). The effect of misinformation, partial information, and guessing on expected multiple-choice test item scores. *Applied Psychological Measurement*, *4*(1), 79-90.

Frary, R. B. (1995). More multiple-choice item writing do's and don'ts. *Practical Assessment, Research & Evaluation*, *4*(11). [http://pareonline.net/getvn.asp?v=4&n=11](https://mail.middlebury.edu/owa/redir.aspx?C=b2q_cSbaX02BPFwkxvdWSXom62UbztAIpG-kUADfgLX623qjim45pT6kahAkCi23obNixX-WpqM.&URL=http%3a%2f%2fpareonline.net%2fgetvn.asp%3fv%3d4%26n%3d11)

Frary, R. B., Tideman, T. N., & Watts, T. M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational and Behavioral Statistics*, *2*(4), 235-256.

Frederick, R. I., & Foster, H. G. (1991). Multiple measures of malingering on a forced-choice test of cognitive ability. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *3*(4), 596-602.

Freedle, R., & Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing*, *16*(1), 2-32.

Friedman, S. & Cook, G. (1995). Is an examinee’s cognitive style related to the impact of answer-changing on multiple-choice tests? *Journal of Experimental Education, 63*(3), 199-213.

Fuhrman, M. (1996). Developing good multiple-choice tests and test questions. *Journal of Geoscience Education*, *44*(4), 379-84.

Geiger, M. (1991a). Changing multiple choice answers: A validation and extension. *College Student Journal, 25*(2), 181-186.

Geiger, M. (1991b). Changing multiple-choice answers: Do students accurately perceive their performance? *The Journal of Experimental Education, 59*(3), 250-257.

Geiger, M. (1996). On the benefits of changing multiple-choice answers: Student perception and performance. *Education, 117*, 108-116.

Green, K. (1981). Item-response changes on multiple-choice tests as a function of test anxiety. *Journal of Experimental Education, 49*(4), 225-228.

Haladyna, T. M., & Downing, S. M. (1989). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, *2*(1), 51-78.

Haladyna, T. M. (2012). *Developing and validating multiple-choice test items*. Routledge.

Haladyna, T. M., & Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied measurement in education*, *2*(1), 37-50.

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item?. *Educational and Psychological Measurement*, *53*(4), 999-1010.

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), 309-333.

Haladyna, T. M., & Shindoll, R. R. (1989). Item shells: A method for writing effective multiple-choice test items. *Evaluation & the Health Professions*, *12*(1), 97-106.

Hambleton, R. K., Roberts, D. M., & Traub, R. E. (1970). A comparison of the reliability and validity of two methods for assessing partial knowledge on a multiple‐choice test. *Journal of Educational Measurement*, *7*(2), 75-82.

Hancock, G. R. (1994). Cognitive complexity and the comparability of multiple-choice and constructed-response test formats. *The Journal of Experimental Education*, *62*(2), 143-157.

Hansen, J. D., & Dexter, L. (1997). Quality multiple-choice test questions: Item-writing guidelines and an analysis of auditing testbanks. *Journal of Education for Business*, *73*(2), 94-97.

Hassmén, P., & Hunt, D. P. (1994). Human self‐assessment in multiple‐choice testing. *Journal of Educational Measurement*, *31*(2), 149-160.

Heim, A. W., & Watts, K. P. (1967). An experiment on multiple-choice versus open-ended answering in a vocabulary test. *British Journal of Educational Psychology*, *37*(3), 339-346.

Helwig, R., Rozek-Tedesco, M. A., Tindal, G., Heath, B., & Almond, P. J. (1999). Reading as an access to mathematics problem solving on multiple-choice tests for sixth-grade students. *The Journal of Educational Research*, *93*(2), 113-125.

Hemmati, F., & Ghaderi, E. (2014). The effect of four formats of multiple-choice questions on the listening comprehension of EFL learners. *Procedia-Social and Behavioral Sciences*, *98*, 637-644.

Hidayati, T., & Yudha, R. P. (2023). Development of higher order thinking skills-based multiple choice test items using Quizizz application to measure the cognitive abilities of early childhood students in the early beginning lasses. *Edumaspul: Jurnal Pendidikan*, *7*(1), 1192-1200.

Hijji, B. M. (2017). Flaws of multiple choice questions in teacher-constructed nursing examinations: A pilot descriptive study. *Journal of Nursing Education*, *56*(8), 490-496.

Horst, P. (1933). The difficulty of a multiple choice test item. *Journal of Educational Psychology*, *24*(3), 229-232.

Imura, H. (2010). Factors affecting listening performance on multiple-choice tests: The effects of stem/option preview and text characteristics. *Language Education & Technology*, *47*, 17-36.

In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing, 26*(2), 219-244. [http://ltj.sagepub.com/content/26/2/219.full.+ html](http://ltj.sagepub.com/content/26/2/219.full.+%20html)

Jones, G. (2021). Designing multiple-choice test items. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 90-101). Routledge.

Karandikar, R. L. (2010). On multiple choice tests and negative marking. *Current Science*, *99*(8), 1042-1045.

Karimi, M., & Biria, R. (2017). Impact of risk taking strategies on male and female EFL learners’ test performance: The case of multiple choice questions. *Theory and Practice in Language Studies*, *7*(10), 892-899. https://www.academypublication.com/issues2/tpls/vol07/10/10.pdf

Kehoe, J. (1995). Writing multiple-choice test items. *Practical Assessment, Research & Evaluation, 4*(9). http://PAREonline.net/getvn.asp?v=4&n=9.

Khan, H. F., Danish, K. F., Awan, A. S., & Anwar, M. (2013). Identification of technical item flaws leads to improvement of the quality of single best multiple choice questions. *Pakistan Journal of Medical Sciences*, *29*(3), 715.

Kruglov, L. P. (1953). Qualitative differences in the vocabulary choices of children as revealed in a multiple-choice test. *Journal of Educational Psychology*, *44*(4), 229-243.

Kulhavy, R. W., & Anderson, R. C. (1972). Delay-retention effect with multiple-choice tests. *Journal of Educational Psychology*, *63*(5), 505-512.

Kumazawa, T. (2016). Factors affecting multiple-choice cloze test score variance: A perspective from generalizability theory. *International Journal of Language Studies*, *10*(1), 15-30.

Lee, H., & Winke, P. (2013). The differences among three-, four-, and five-option-item formats in the context of a high-stakes English-language listening test. *Language Testing, 30*(1), 99-123.

Lehrl, S., Triebig, G., & Fischer, B. (1995). Multiple choice vocabulary test MWT as a valid and short test to estimate premorbid intelligence. *Acta Neurologica Scandinavica*, *91*(5), 335-345.

Lesage, E., Valcke, M., & Sabbe, E. (2013). Scoring methods for multiple choice assessment in higher education–Is it still a matter of number right scoring or negative marking?. *Studies in Educational Evaluation*, *39*(3), 188-193.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational and Behavioral Statistics*, *4*(4), 269-290.

Lions, S., Monsalve, C., Dartnell, P., Godoy, M. I., Córdova, N., Jiménez, D., ... & Lemarié, J. (2021). The position of distractors in multiple-choice test items: The strongest precede the weakest. *Frontiers in Education,* 6, 731-763.

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science*, *23*(11), 1337-1344.

Lord, F. M. (1952). The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika*, *17*(2), 181-194.

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple‐choice, constructed response, and examinee‐selected items on two achievement tests. *Journal of Educational Measurement*, *31*(3), 234-250.

Lunrasri, Y., Tangdhanakanond, K., & Pasiphol, S. (2022). Item analysis of multiple-choice reading literacy instruments using item response theory. *Journal of Education Naresuan University, 24*(4), 61-72.

Marsh, E. J., Roediger, H. L., Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review*, *14*(2), 194-199.

Martínez, R. J., Moreno, R., Martín, I., & Trigo, M. E. (2009). Evaluation of five guidelines for option development in multiple-choice item-writing. *Psicothema*, *21*(2), 326-330.

Mason, V. (1984). Using multiple-choice tests to promote homogeneity of class ability levels in large EGP and ESP programs. *System, 12*(3), 263-271.

Mason, V. (1992). A good word for multiple-choice tests. *CATESOL Journal, 5*(2), 29-44.

Masters, J. C., Hulsmeyer, B. S., Pike, M. E., Leichty, K., Miller, M. T., & Verst, A. L. (2001). Assessment of multiple-choice questions in selected test banks accompanying text books used in nursing education. *The Journal of Nursing Education*, *40*(1), 25-32.

McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: A literature review. *Medical Teacher*, *26*(8), 709-712.

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, *4*(2), 142-154.

Mehrens, W.A. & Lehman, I.J. (1978). *Measurement and evaluation in education and psychology* (2nd ed.). Holt, Rinehart and Winston.

Midway, S., Robertson, M., Flinn, S., & Kaller, M. (2020). Comparing multiple comparisons: Practical guidance for choosing the best multiple comparisons test. *PeerJ*, *8*, e10387.

Miklas, M. (2021). One of these does not belong: Creating interesting multiple-choice questions for teaching and testing. In *English Teaching Forum, 59*(1), 31-35.

Mitkov, R., An Ha, L., & Karamanis, N. (2006). A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering*, *12*(02), 177-194.

Morrison, S., & Free, K. W. (2001). Writing multiple-choice test items that promote and measure critical thinking. *Journal of Nursing Education*, *40*(1), 17-24.

Nedeau-Cayo, R., Laughlin, D., Rus, L., & Hall, J. (2013). Assessment of item-writing flaws in multiple-choice questions. *Journal for Nurses in Professional Development*, *29*(2), 52-57.

Nevo, N. (1989). Test-taking strategies on a multiple-choice test of reading comprehension. *Language Testing*, *6*(2), 199-215.

Nicol, D. (2007). E‐assessment by design: Using multiple‐choice tests to good effect. *Journal of Further and Higher Education*, *31*(1), 53-64.

Norris, S. P. (2009). Informal reasoning assessment: Using verbal reports of thinking to improve multiple-choice test validity. In J. F. Voss, D. N. Perkins, & J. W. Segal (Eds.), *Informal reasoning and education* (pp. 451-471). Routledge.

O’Grady, S. (2023). Adapting multiple-choice comprehension question formats in a test of second language listening comprehension. *Language Teaching Research*, *27*(6), 1431-1455.

Oller, J.W., Jr. (1979). *Language tests at school.* Longman.

Özdemir, A. (2021). Multiple choice item writing and analysis (A framework for action). In K. Büyükkarcı & A. Önal (Eds.), Essentials of applied linguistics and foreign language teaching: 21st century skills and classroom applications (pp. 190-208). ISRES Publishing

Öztürk, M. (2007). Multiple-choice test items of foreign language vocabulary. *Uludağ Üniversitesi Eğitim Fakültesi Dergisi*, *20*(2), 399-426.

Pamphlett, R., & Farnill, D. (1995). Effect of anxiety on performance in multiple choice examination. *Medical Education*, *29*(4), 297-302.

Papenberg, M., Diedenhofen, B., & Musch, J. (2021). An experimental validation of sequential multiple-choice tests. *The Journal of Experimental Education*, *89*(2), 402-421.

Paxton, M. (2000). A linguistic perspective on multiple-choice questioning. *Assessment & Evaluation in Higher Education, 25*(2), 109-119.

Pezeshkpour, P., & Hruschka, E. (2023). Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*.

Pressley, M., & Ghatala, E. S. (1988). Delusions about performance on multiple-choice comprehension tests. *Reading Research Quarterly*, 454-464.

Pressley, M., Ghatala, E. S., Woloshyn, V., & Pirie, J. (1990). Sometimes adults miss the main ideas and do not realize it: Confidence in responses to short-answer and multiple-choice comprehension questions. *Reading Research Quarterly*, 232-249.

Pyrczak, F. (1972). Objective evaluation of the quality of multiple-choice test items designed to measure comprehension of reading passages. *Reading Research Quarterly, 8*(1), 62-71.

Rahimirad, M. (2014). The impact of metacognitive strategy instruction on the listening performance of university students. *Procedia-Social and Behavioral Sciences*, *98*, 1485-1491.

Rankin, E. F., & Culhane, J. W. (1969). Comparable cloze and multiple-choice comprehension test scores. *Journal of Reading*, *13*(3), 193-198.

Rhind, S. M., & Pettigrew, G. W. (2012). Peer generation of multiple-choice questions: student engagement and experiences. *Journal of Veterinary Medical Education*, *39*(4), 375-379.

Ruohoniemi, M., & Lindblom‐Ylänne, S. (2009). Students' experiences concerning course workload and factors enhancing and impeding their learning–a useful resource for quality enhancement in teaching and curriculum planning. *International Journal for Academic Development*, *14*(1), 69-81.

Rodriguez, M. C. (2003). Construct equivalence of multiple‐choice and constructed‐response items: A random effects synthesis of correlations. *Journal of Educational Measurement*, *40*(2), 163-184.

Rodriguez, M. C. (2005). Three options are optimal for multiple‐choice items: A meta‐analysis of 80 years of research. *Educational Measurement: Issues and Practice*, *24*(2), 3-13.

Roediger III, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 1155.

Roid, G.H., & Haladyna, T.M. (1980). The emergence of an item-writing technology. *Review of Educational Research, 50*(2), 293-314.

Rosenthal, R., & Rubin, D. B. (1989). Effect size estimation for one-sample multiple-choice-type data: Design, analysis, and meta-analysis. *Psychological Bulletin*, *106*(2), 332-337.

Rubin, M. (2021). When to adjust alpha during multiple testing: a consideration of disjunction, conjunction, and individual testing. *Synthese*, *199*(3), 10969-11000.

Rupp, A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing, 23*(4), 441-474.

Saeedi, M. (2016). Construct validity of multiple-choice cloze test and cloze-elide test in testing used for reading comprehension among Iranian EFL Learners. *International Research Journal of Applied Basic Science*, *10*(6), 632-637.

Savelka, J., Agarwal, A., Bogart, C., & Sakr, M. (2023). Large language models (gpt) struggle to answer multiple-choice questions about code. *arXiv preprint arXiv:2303.08033*.

Schultheis, N. M. (1998). Writing cognitive educational objectives and multiple-choice test questions. *American Journal of Health-system Pharmacy*, *55*(22), 2397-2401.

Scouller, K. M. (1997). Students' perceptions of three assessment methods: Assignment essay, multiple choice question examination, short answer examination. *Research and Development in Higher Education*, *20*, 646-653.

Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education*, *35*(4), 453-472.

Scouller, K. M., & Prosser, M. (1994). Students' experiences in studying for multiple choice question examinations. *Studies in Higher Education*, *19*(3), 267-279.

Shin, J., Guo, Q., & Gierl, M. J. (2019). Multiple-choice item distractor development using topic modeling approaches. *Frontiers in Psychology*, *10*, 825. https://www.frontiersin.org/articles/10.3389/fpsyg.2019.00825/full

Shirzadi, D., & Amerian, M. (2020). Washback effects of multiple-choice, cloze and metalinguistic tests on EFL students writing. *Indonesian Journal of Applied Linguistics*, *9*(3), 536-544.

Shizuka, T., Takeuchi, O., Yashima, T. & Yoshizawa, Y. (2006). A comparison of 3 and 4 option English tests for university entrance selection purposes in Japan. *Language Testing*, *23*(1), 35-57.

Smith, J. K. (1982). Converging on correct answers: A peculiarity of multiple-choice items. *Journal of Educational Measurement, 19*(3), 211-220.

Spaan, M. (2007). Evolution of a test item. *Language Assessment Quarterly, 4*(3), 279-293. [http://www.tandfonline.com/doi/pdf/10.1080/15434300701462937](https://mail.middlebury.edu/owa/redir.aspx?C=gr0BLvrcFki4pdfj_1qh6sVKqt1UztEIp1unuG4vGM2L4seXpMkli5TyCnAgO13_Xl0dgDHSwW8.&URL=http%3a%2f%2fwww.tandfonline.com%2fdoi%2fpdf%2f10.1080%2f15434300701462937)

Spolsky, B. (1986). A multiple choice for language testers.*Language Testing, 3*, 147-158.

Steinberg, R. N., & Sabella, M. S. (1997). Performance on multiple-choice diagnostics and complementary exam problems. *Physics Teacher*, *35*, 150-155.

Stewart, J. (2014). Do multiple-choice options inflate estimates of vocabulary size on the VST.  *Language Assessment Quarterly, 11*(3), 271-282. [http://www.tandfonline.com/doi/pdf/10.1080/15434303.2014.922977](https://mail.middlebury.edu/owa/redir.aspx?C=gr0BLvrcFki4pdfj_1qh6sVKqt1UztEIp1unuG4vGM2L4seXpMkli5TyCnAgO13_Xl0dgDHSwW8.&URL=http%3a%2f%2fwww.tandfonline.com%2fdoi%2fpdf%2f10.1080%2f15434303.2014.922977)

Stoeckel, T., & Sakigara, T. (2018). A serial multiple-choice format designed to reduce overstimulation of meaning-recall knowledge on the Vocabulary Size Test. *TESOL Quarterly, 52*(4), 1050-1062.

Stanger-Hall, K. F. (2012). Multiple-choice exams: an obstacle for higher-level thinking in introductory science classes. *CBE—Life Sciences Education*, *11*(3), 294-306.

Tamir, P. (1971). An alternative approach to the construction of multiple choice test items. *Journal of Biological Education*, *5*(6), 305-307.

Tamir, P. (1993). Positive and negative multiple choice items: How different are they? *Studies in Educational Evaluation*, *19*(3), 311-325.

Tarrant, M., Knierim, A., Hayes, S. K., & Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education in Practice*, *6*(6), 354-363.

Tarrant, M., & Ware, J. (2008). Impact of item‐writing flaws in multiple‐choice questions on student achievement in high‐stakes nursing assessments. *Medical Education*, *42*(2), 198-206.

Tarrant, M., Ware, J., & Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. *BMC medical education*, *9*(1), 40.

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika*, *49*(4), 501-519.

Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple‐choice models: The distractors are also part of the item. *Journal of Educational Measurement*, *26*(2), 161-176.

Thissen, D., Wainer, H., & Wang, X. B. (1994). Are tests comprising both multiple‐choice and free‐response items necessarily less unidimensional than multiple‐choice tests? An analysis of two tests. *Journal of Educational Measurement*, *31*(2), 113-123.

Tiara, E. (2019). Effectiveness of test design between of cloze test and multiple choice test for reading comprehension. *Journal of Education, Linguistics, Literature and Language Teaching*, *2*(02), 53-68.

Tiara, E., Chairuddin, C., Makhroji, M., Zulida, E., Fadlia, F., & Rahmiati, R. (2019). Effectiveness of test design between of cloze test and multiple choice test for reading comprehension. *Journal of Education, Linguistics, Literature and Language Teaching*, *2*(02), 53-68.

Tinkelman, S. N. (1968). Checklist for reviewing local school tests. In N. E. Gronlund (Ed.), *Readings in measurement and evaluation* (pp. 103-108). McMillan.

Toksöz, S., & Ertunç, A. (2017). Item analysis of a multiple-choice exam. *Advances in Language and Literary Studies*, *8*(6), 141-146.

Tovani, C. (2023). *So what do they really know?: Assessment that informs teaching and learning*. Routledge.

Traub, R. E., & Fisher, C. W. (1977). On the equivalence of constructed-response and multiple-choice tests. *Applied Psychological Measurement*, *1*(3), 355-369.

Treagust, D. (1986). Evaluating students' misconceptions by means of diagnostic multiple choice items. *Research in Science Education*, *16*(1), 199-207.

Votaw, D. F. (1936). The effect of do-not-guess directions upon the validity of true-false or multiple choice tests. *Journal of Educational Psychology*, *27*(9), 698-703.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education*, *6*(2), 103-118.

Ward, W. C. (1982). A comparison of free-response and multiple-choice forms of verbal aptitude tests. *Applied Psychological Measurement*, *6*(1), 1-11.

Webb, E. M., Phuong, J. S., & Naeger, D. M. (2015). Does educator training or experience affect the quality of multiple-choice questions?. *Academic Radiology*, *22*(10), 1317-1322.

Wesman, A.G. (1971). Writing the test item. In R.L. Thorndike (Ed.) *Educational measurement* (1st ed., pp. 99-111). American Council on Education.

Whitley, T. W. (1979). Some common flaws in multiple choice exam questions. *Nursing Outlook*, *27*(7), 466-69.

Wilhite, S. C. (1986). The relationship of headings, questions, and locus of control to multiple-choice test performance. *Journal of Literacy Research*, *18*(1), 23-40.

Willey, C. F. (1960). The three-decision multiple-choice test: A method of increasing the sensitivity of the multiple-choice item. *Psychological Reports*, *7*(3), 475-477.

Wu, A. D., Park, M., & Hu, S. F. (2021). Gender fairness in immigration language testing: A study of differential options functioning on the CELPIP-G reading multiple-choice questions. *International Journal of Quantitative Research in Education*, *5*(3), 244-267.

Yanagawa, K., & Green, A. (2008). To show or not to show: The effects of item stems and answer options on performance on a multiple-choice listening comprehension test. *System*, *36*(1), 107-122.

Yi'an, W. (1998). What do tests of listening comprehension test? A retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing*, *15*(1), 21-44.

Yonker, J. E. (2011). The relationship of deep and surface study approaches on factual and applied test‐bank multiple‐choice question performance. *Assessment & Evaluation in Higher Education*, *36*(6), 673-686.

Zandi, B., Roushan, B., Vakilifard, A., & Golpur, L. (2014). Multiple-choice listening tests: Spoken choices or written choices?. *Journal of Teaching Persian to Speakers of Other Languages*, *3*(7), 47-68.

Zeidner, M. (1987). Essay versus multiple-choice type classroom exams: The student’s perspective. *The Journal of Educational Research*, *80*(6), 352-358.

Zhang, Z., Lei, L., Wu, L., Sun, R., Huang, Y., Long, C., ... & Huang, M. (2023). Safetybench: Evaluating the safety of large language models with multiple choice questions. *arXiv preprint arXiv:2309.07045*.

Zimmerman, D. W., & Williams, R. H. (1965). Chance success due to guessing and non-independence of true scores and error scores in multiple-choice tests: Computer trials with prepared distributions. *Psychological Reports*, *17*(1), 159-165.