Oregon Science Assessment Peer Review Resubmission Evidence

Critical Element 3.3:
Dimensionality Study

June 2023

OREGON
DEPARTMENT OF EDUCATION
Oregon achieves . . . together!

Introduction

The Oregon Department of Education's (ODE) Next Generation Science Standards (NGSS) assessment consists of multiple conceptions of dimensionality. In an effort to avoid misconceptions with respect to the internal structure of ODE's NGSS assessment, the purpose of this study is to examine and discuss each conception of dimensionality with the aim of (1) clearly articulating the dimensional structure of the assessment and (2) satisfying the requirements of critical element 3.3 (i.e., *validity based on internal structure*). The conceptions of dimensionality are the following: multidimensional items, item types with nuisance dimensions, test assembly design (i.e., content balancing on three domains), and the estimation of a unidimensional scale.

Conceptions of Dimensionality

1. *Multidimensional items*: Each item on ODE's NGSS assessment corresponds to a performance expectation (PE), and each PE consists of three dimensions (i.e., science and engineering practices [SEPs], crosscutting concepts [CCCs], and disciplinary core ideas [DCIs]). This suggests items are multidimensional. Each item's PE requires a student to integrate the SEP, CCC, and DCI in order to identify and interpret evidence, engage in scientific reasoning, make sense of phenomena, and address problem(s) (Achieve, 2018). While the items are overtly multidimensional, a student's response is unidimensional because the response is the integration of the three dimensions. This conception of dimensionality is distinct from other statewide summative assessments. For instance, the items within ODE's English language arts (ELA) assessment are explicitly unidimensional and correspond to a single target[1].

   We realize and understand that one might suppose ODE's NGSS assessment has a three-dimensional internal structure given the dimensionality of the items. However, it is impossible to disentangle the dimensions for the purposes of estimating a student's unique SEP, CCC, or DCI performance. This is because items load on all three dimensions given that PEs require students to integrate the SEPs, CCCs, and DCIs. For instance, if we were to estimate a score representing the SEP dimension, this score would invariably include information from the CCC and DCI dimensions. It is not clear how one would use or interpret separate SEP, CCC, or DCI scores knowing they also include information from the other dimensions. Therefore, despite the multidimensional items, we expect ODE's

[1] It is important to note that items from ELA or any content area assessment may implicitly require the integration of other dimensions (e.g., reading and writing) which would certainly raise questions and concerns about the internal structure of the assessment.

NGSS assessment to have a unidimensional internal structure measuring a student's integration of the three dimensions (see also Kaldaras, Akaeze, & Krajcik, 2021a; 2021b).

2. *Nuisance dimensions*: ODE's NGSS items are testlets consisting of one or more assertions, and assertions are the scored student interactions within each item. Moreover, there are two item types within ODE's NGSS assessment (i.e., standalone and cluster items). Standalone items are testlets with three or fewer assertions, and cluster items are testlets with four or more assertions. Note that assertions within standalone and cluster items share a common stimulus; thus, they are locally dependent. This local dependence results in the presence of a nuisance dimension for each item and, if left alone, will bias item parameters, underestimate standard errors, and overestimate the marginal reliability. ODE, partner states, and Cambium Assessment use a Rasch testlet model to account for the nuisance dimensions as part of item calibration and score estimation. The use of the Rasch testlet model is an acceptable method to account for dimensionality resulting from local item dependence (Jiao, Wang, & He, 2013; Wang & Wilson, 2005). Lastly, it is important to acknowledge that other statewide summative assessments (e.g., English language arts) typically ignore local item dependence. That is, those assessments contain groups of items sharing a common stimulus (e.g., reading passage) and ignore the impact of nuisance dimensions on the item parameter estimates, standard errors, and the marginal reliability.

3. *Test assembly design*: ODE's NGSS assessment used a linear-on-the-fly test (LOFT) assembly design to deliver items to students during the 2018-19 school year. The LOFT assembly design randomly selected stand-alone and cluster items for each student in order to meet the content blueprint requirements (e.g., a specific number of stand-alone and cluster items for each domain—earth and space science [ESS], life science [LS], and physical science [PS]). The rationale to balance content according to domains is simply convenience. As mentioned above, all NGSS items correspond to a three-dimensional PE. A nice feature of the DCI dimension is every item associates with one of three domains (i.e., ESS, LS, and PS), and the domains have a robust intersection with SEPs and CCCs. Balancing content by domain is a convenient way to ensure the LOFT assembly design delivers a reasonably similar and representative set of PEs to each student.

   Because of the convenience of using domains as part of the test assembly design, ODE also estimates and reports ESS, LS, and PS domain scores for each student. This is certainly a controversial and contradictory decision because it leads one to assume the underlying structure of ODE's NGSS assessment is multidimensional (Feinberg & Jurich, 2017;

Haberman, 2008). Moreover, it is a clear indication that the scoring and reporting structures are not consistent with the NGSS. Our rationale to report domain scores is to comply with clauses (x) and (xii) of ESEA Section 1111(b)(2)(B). ODE, along with most states, believe that reporting domain scores (i.e., subscores) is the only way to comply with the diagnostic requirements of clauses (x) and (xii) (Marion & Briggs, 2022). In lieu of reporting domain scores, ODE's preference would be to identify a methodology to support the reporting of meaningful achievement information to parents or guardians, educators, and school administrators. ODE recently had a conversation with colleagues from the U.S. Department of Education about compliance with clauses (x) and (xii), and learned that the use of domain scores is not a requirement or expectation of those clauses. The U.S Department of Education invited ODE to explore innovative approaches to satisfy clauses (x) and (xii) while meeting the reporting needs of parents or guardians, education partners, and community partners.

4. *Unidimensional scale*: ODE's June 2020 peer review submission to the U.S. Department of Education for critical element 3.3 included the results from three dimensionality analyses (i.e., principal components analysis [PCA], confirmatory factor analysis [CFA], and a confirmatory DETECT analysis). The PCA and CFA were preliminary analyses with the intention of signaling the plausibility of unidimensionality. Using the domain scores as the manifest variables, the PCA found only one component with an eigenvalue greater than one (i.e., Kaiser's rule). The CFA, estimating a single latent NGSS construct via the domain scores, corroborated the plausibility of unidimensionality. While the PCA and CFA were preliminary models, their findings suggested a unidimensional internal structure similar to what Kaldaras, Akaeze, and Krajcik (2021b) found in their study of a NGSS-aligned assessment.

   The confirmatory DETECT analysis using the SIRT R package (Robitzsch, 2022) was a more robust examination of dimensionality. Specifically, the aim was to indicate whether a matrix of item responses had an underlying unidimensional structure by examining the covariances between pairs of items conditional on the observed student ability estimate (Zhang, 2013; Zhang & Stout, 1999; Stout, 1996). The confirmatory DETECT analysis from our June 2020 peer review submission focused only on the influence of the nuisance dimensions (i.e., multidimensionality due to items having a common stimulus). The findings matched our expectations concerning the nuisance dimensions and their influence. That is, given the testlet design of ODE's NGSS assessment, we expected to observe a range of dimensionality from unidimensional to weakly multidimensional (despite ODE's use of a Rasch testlet model).

Table 1 below displays the results from a follow-up confirmatory DETECT analysis focusing on two dimensions (i.e., item type and domain) using the same data as the analysis from our June 2020 peer review submission (i.e., Oregon's 2018-19 NGSS assessment). The aim of this follow-up analysis is to examine the conditional covariances of item pairs according to the item type and domain. Note that item type refers to standalone and cluster item types, and domain refers to the ESS, LS, and PS domains. We did not include the three NGSS dimensions (i.e., SEP, CCC, and DCI) because, as we mentioned previously, each NGSS item loads on a specific SEP, CCC, and DCI combination. It is unclear how we would examine the influence of the NGSS dimensions unless we view the intersection of each SEP, CCC, and DCI as a dimension; however, this would mean each PE is a unique dimension. We are not confident this would be a fruitful dimensionality study given the number of PEs in each grade (i.e., 42 in 5th, 55 in 8th, and 67 in 11th). Nonetheless, findings from the follow-up confirmatory DETECT analysis suggest a unidimensional internal structure with respect to the item types and domains as all dimensionality statistics meet the unidimensionality criteria. That is, ODE's NGSS assessment has a unidimensional structure and does not measure separate dimensions by item type or domain. This is a robust confirmation of the findings from the PCA and CFA (and the study by Kaldaras, Akaeze, and Krajcik [2021b]); moreover, it also suggests ODE's reporting of domains is not meaningful as they convey redundant information.

Table 1.
Confirmatory DETECT analysis by grade and dimension.

Grade and Dimension	Dimensionality Statistics		
	DETECT	Approximate Simple Structure Index (ASSI)	Ratio
5th Grade			
Item Type	0.058	0.052	0.040
Domain	0.178	0.124	0.123
8th Grade			
Item Type	-0.070	0.012	-0.050
Domain	0.104	0.088	0.073
11th Grade			
Item Type	-0.043	-0.041	-0.044
Domain	0.151	0.133	0.154

Note. The dimensionality statistics are weighted values (i.e., weighted by the sample size of item pairs). In general, weighted values are larger than unweighted values. Unidimensionality criteria is < 0.20 for DETECT, < 0.25 for ASSI, and < 0.36 for the ratio statistic (Robitzsch, 2022). Negative values for each dimensionality statistic refers to a clear unidimensional structure.

References

Achieve. (2018). *Criteria for procuring and evaluating high-quality and aligned summative science assessments (version 1.0).* Washington, D.C.: Achieve. Retrieved on June 14, 2023, from https://www.nextgenscience.org/sites/default/files/Criteria03202018.pdf

Feinberg, R. A., & Jurich, D. P. (2017). Guidelines for interpreting and reporting subscores. *Educational Measurement: Issues and Practice, 36*(1), 5-13. https://doi.org/10.1111/emip.12142

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics, 33*(2), 204-229. https://doi.org/10.3102/1076998607302636

Jiao, H., Wang, S., & He, W. (2013). Estimation methods for one-parameter testlet models. *Journal of Educational Measurement, 50*(2), 186-203. https://doi.org/10.1111/jedm.12010

Kaldaras, L., Akaeze, H., & Krajcik, J. (2021a). Developing and validating Next Generation Science Standards-aligned learning progression to track three-dimensional learning of electrical interactions in high school physical science. *Journal of Research in Science Teaching, 58*(4), 589-618. https://doi.org/10.1002/tea.21672

Kaldaras, L., Akaeze, H., & Krajcik, J. (2021b). A methodology for determining and validating latent factor dimensionality of complex multi-factor science constructs measuring knowledge-in-use. *Educational Assessment, 26*(4), 241-263. https://doi.org/10.1080/10627197.2021.1971966

Marion, S., & Briggs, D. (2022, July 22). Just give us a little: Please make one small change in federal testing law to yield big improvements. *The National Center for the Improvement of Educational Assessment*. https://www.nciea.org/blog/just-give-us-a-little/

Robitzsch, A. (2022). *sirt: Supplementary item response theory models* (3.12-66). https://CRAN.R-project.org/package=sirt

Stout, W. et al. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 20*(4), 331-354. https://doi.org/10.1177/014662169602000403

Wang, W. C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement, 29*(2), 126-149. https://doi.org/10.1177/0146621604271053

Zhang, J. (2013). A procedure for dimensionality analyses of response data from various test designs. *Psychometrika, 78*(1), 37-58. https://doi.org/10.1007/s11336-012-9287-z

Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*(2), 213-249. https://doi.org/10.1007/BF02294536