doi: 10.21277/se.v1i34.253

# BENDRŲJŲ GAMTOS PAŽINIMO IR SOCIALINIŲ MOKSLŲ MOKYMOSI REZULTATŲ VERTINIMO VALIDUMAS

Paul Mooney
Specialiojo ugdymo programos
Luizianos valstybinis universitetas
213 Peabody Hall, JAV

Renée E. Lastrapes
Edukologijos fakultetas, Hiustono universitetas
UH-Clear Lake Box 55
2700 Bay Area Blvd. Houston, Texas 77058, JAV

Amanda M. Marcotte
Masačusetso universitetas, Amherstas
S154 Furcolo
Amherst, MA 01003, JAV

Amy Matthews, B. S.
Pradinių klasių gamtos pažinimo mokytoja
Luizianos valstybinis universitetas
13046 LA-73, Geismar, LA 70734, JAV

**Anotacija**

Šis tyrimas papildo tyrimo validavimo rezultatus, gautus atlikus dalyko mokymosi struktūruotą formuojamąjį vertinimą, kuris buvo vykdomas internetu. Tyrimas kartu vertino ir galimybę įtraukti papildomus kintamuosius, tokius kaip teksto suvokimo pažymint teisingus ar klaidingus teiginius metodą bei rašytinį atpasakojimą, tam, kad būtų galima paaiškinti mokinių pasiekimų skirtumus gamtos pažinimo ir socialinių mokslų srityse. Tyrime dalyvavo penktos klasės mokiniai (N = 51), lankantys valstybinę pradinę mokyklą pietrytinėje JAV dalyje. Trys kintamieji – prediktoriai (t. y. turinio suvokimas, klaidingų ar teisingų teiginių žymėjimo metodas ir rašytinis atpasakojimas) koreliavo su testo rezultatais, gautais atlikus nacionalinį reprezentuojamąjį standartizuotą pasiekimų testą (Stanfordo pasiekimų testo dešimtoji sutrumpinta internetinė versija) ir valstijos atsiskaitomąjį testą. Pirsono (Pearson) koreliacija tarp turinio suvokimo ir Stanfordo gamtos pažinimo (r = ,55) ir socialinių mokslų (r = ,63) sričių testų buvo vidu-

tiniškai stipri ir dydžiu panaši į kitas koreliacijas akademinės kalbos tyrimo atvejais. Turinio suvokimo koreliacija buvo didesnė nei nustatytosios tarp standartizuotų testų ir teisingų ar klaidingų teiginių žymėjimo metodo bei rašytinio atpasakojimo. Panašumų analizė rodo, kad tiek turinio suvokimas, tiek teisingų ar klaidingų teiginių žymėjimo metodas aiškinamuosius modelius papildė unikaliais skirtumais. Straipsnyje taip pat aptariami tyrimo ribotumai.

**Esminiai žodžiai:** *struktūruotas formuojamasis vertinimas, bendras rezultatų vertinimas, turinio mokymasis.*

Ilgą laiką struktūruotas formuojamasis vertinimas buvo specialiojo ugdymo dalis tiek profesine (praktine), tiek ir teorine prasme (Jenkins ir Fuchs, 2012). Nuo dr. Stanley'o Deno išplėtoto ir pasiūlyto mokymosi programa grįsto vertinimo laikų (Deno, 1985) specialieji pedagogai turėjo galimybę stebėti mokinių pažangą, siekdami individualizuotų ugdymo programos tikslų. Mokymosi programa (angl. *Curriculum*) grįstas mokinių pasiekimų vertinimas buvo įdiegtas pirmiausia pradinėse klasėse, pradėjus ugdyti mokinių skaitymo įgūdžius. Šios vertinimo procedūros buvo taikomos ir vertinant mokinių matematikos, taisyklingos rašybos, rašymo žinias ir įgūdžius. Tačiau tai, kaip mokiniams sekasi mokytis ir suprasti socialinių mokslų ir gamtos pažinimo turinį, nebuvo vertinama. Būtent tai ir vertina šis tyrimas.

Struktūrinio formuojamojo vertinimo veiksmingumas, siekiant užfiksuoti mokinių pasiekimus ir jų pažangą socialinių mokslų ir gamtos pažinimo mokymosi srityse, vis dar yra mažai tyrinėtas. Espin ir Foegen (1996) domėjosi, ar įmanoma, kad skaitymo vertinimas, pavyzdžiui, skaitymo balsu sklandumas, padėtų mokiniams geriau suvokti dalyko turinį. Tyrimo rezultatai parodė, kad dalyko turinio įsisavinimas stipriau koreliuoja su jo pateikimu rašytine nei sakytine forma. Iš čia išsirutuliojo daugiau mokymosi turiniu grįstų vertinimų, įskaitant ir turinio vertinimo validumo testus (Mooney, McCarter, Russo, Blackwood, 2013), internetu administruojamą struktūruotą formuojamąjį vertinimą, kuris ir yra šio tyrimo objektas. Straipsnyje aprašomas mokinių turinio suvokimo įsisavinimas ir pagrindžiami du tyrimo klausimai:

1. Koreliacija su nacionaliniais norminiais standartizuotais pasiekimų ir valstijos atsiskaitomaisiais testais gamtos pažinimo ir socialinių mokslų dalyko turinio suvokimo, taikant teisingų ar klaidingų teiginių patvirtinimo metodą ir rašytinį atpasakojimą, atvejais.

2. Tyrimo validumas, papildžius skaitomo teksto supratimo vertinimą teisingų ar klaidingų teiginių žymėjimo metodu ir (ar) rašytinio atpasakojimo turinio suvokimo vertinimu.

## Turinio suvokimo vertinimas

Turinio suvokimo vertinimas buvo sukurtas taikant mokymo programos atrankos būdą (Fuchs, 2004), t. y. tyrimai buvo atliekami atrenkant mokymo tekstus tam tikros klasės mokymo programai. Mokymosi turinio vertinimo tyrimas apima mokinių skaitomus svarbiausius gamtos pažinimo ir socialinių mokslų srities tekstus, žodyną, gebėjimą pasirinkti teisingą atsakymą iš pateikto sąvokų sąrašo. Mokiniams pateikiama 20 klausimų, atsakymams skiriama iki 5 min. laiko.

Naudojamas akademinis žodynas padėjo geriau suprasti dalyko turinį (Alexander, n. d.) ir dalykinę leksiką (žodžius, frazes ir sąvokas). Dalyko mokymosi sėkmę dažnai lemia tinkamo ir suprantamo akademinio žodyno vartojimas. Akademinės kalbos vartojimas stipriau koreliuoja su turinio kriterijumi, siejant skaitymo balsu sklandumą ir atpasakojimo rašymo vertinimą (Espin ir Foegen, 1996; Mooney, McCarter, Schraven, Callicoatte, 2013).

Turinio suvokimo vertinimas yra internetinė žodyno atitikties adaptacija (Espin ir Deno, 1994–1995). Tyrimas skirtas vieno tyrimo rezultato techniniams klausimams analizuoti. Mooney, McCarter, Russo ir kt. (2013) įvertino validumo kriterijų 20-ies gamtos pažinimo tyrimų rinkiniui, siedami jį su valstijos atsiskaitomuoju testu, kurį atliko penktos klasės daugiausia gerai besimokantys mokiniai. Rezultatai parodė 20-ies tyrimų (r = ,36–,55) vidutinę koreliaciją. Mooney, McCarter, Russo ir Blackwood (2014) išplėtė šio kriterijaus validumo rezultatus, atskleisdami vidutiniškai stiprią koreliaciją (,67) tarp turinio suvokimo vertinimo socialinių mokslų mokymosi tyrime ir valstijos turinio testo. Socialinių mokslų turinio mokymosi koreliacija buvo didesnė nei gamtos pažinimo ir panaši į ankstesnius žodyno atitikties rezultatus (rs nuo ,64 iki ,70; Espin, Shin, Busch, 2005; Mooney, McCarter, Schraven ir kt., 2013).

## Tyrimo pagrindimas

Šiame tyrime numatytas vertinimas buvo sukurtas kaip bendrasis mokymosi rezultatų vertinimas ir pritaikytas specialiai šiam tikslui. Bendrasis rezultatų vertinimas yra vienas iš mokymui aktualių vertinimo modelių (Fuchs ir Deno, 1991). Kaip alternatyva įgūdžių įvaldymui matuoti, pirminis tikslas buvo pa-

rengti testus, kurie (a) parodytų patikimumą ir validumą bei (b) padėtų mokytojams sudaryti geresnes ugdymo programas ir vertinti jų sėkmę. Gauti rezultatai papildo mokslinę literatūrą, analizuojančią į turinį orientuotą struktūruotą formuojamąjį vertinimą, taip pat ir tyrimus, skirtus įvertinti mokymosi veiklas ir žodyno pažangą (pvz., Espin, Lembke, Hampton, Seo, Zukowski, 2013; Mooney, McCarter, Schraven ir kt., 2013).

Bendrasis dalykinės kalbos rezultatų vertinimas, įskaitant turinio suvokimą, žodyno atitiktį ir vartojamą žodyną (Vannest, Parker, Dyer, 2011), padeda numatyti mokymosi procesą ir galimus pasiekimus. Turinio suvokimo vertinimas buvo skirtas mokytojams kaip gamtos pažinimo ir socialinių mokslų dalykų vertinimo priemonė aukštesnėse pradinėse ir vidurinėse klasėse. Mokytojai dalykinės kalbos, kaip priklausomo kintamojo (Bloom, 1980), mokymąsi labiau sieja su prasčiau besimokančiais mokiniais. Vis dėlto būtų naudinga atlikti daugiau ir įvairesnių šios srities tyrimų. Teisingų ar klaidingų teiginių žymėjimo metodas ir rašytinis atpasakojimas buvo vertinami kaip teksto supratimą atskleidžiantys bendrieji rodikliai (Marcotte ir Hintze, 2009). Ypač skaitymo atveju pastebėta, kad bendrųjų rodiklių vertinimas buvo veiksmingas įrankis, padedantis numatyti mokinio pasiekimus ir suteikti informacijos mokytojams, priimantiems su mokymo turiniu susijusius sprendimus (Stecker, Fuchs, Fuchs, 2005; Wallace, Espin, McMaster, Deno, Foegen, 2007). Be to, rezultatų vertinimas buvo naudojamas kaip instrumentas įdiegiant atsako į intervenciją schemas.

Anksčiau minėti tyrimo klausimai yra susiję su tuo, ką Fuchs (2004) įvardijo kaip pirmosios pakopos tyrimo rezultatus, o Deno ir Fuchs (1987) nurodė, jog tai jiems leido parinkti adekvačius klausimus instrumentinei matricai parengti. Pirmasis klausimas buvo orientuotas į poreikį išplėsti turinio supratimo vertinimo tyrimo validumą. Šis tyrimas palygina matavimo validumo rezultatus su valstijos atsiskaitomuoju gamtos pažinimo ir socialinių mokslų mokymosi testu (Mooney, McCarter, Russo ir kt., 2013, 2014). Pirmasis tyrimo klausimas išplėtė kriterijaus validumo įrodymą, palygindamas turinio supratimo vertinimo rezultatus su gautaisiais atlikus standartizuotą žinių vertinimą. Apibendrinti nurodytus tyrimus sudėtinga, nes iki šiol turinio supratimo vertinimas buvo lyginamas su specialiais valstijos lygmens atsiskaitomaisiais testais. Šis tyrimas pirmą kartą leido palyginti turinio supratimo vertinimo rezultatus su standartizuoto nacionalinio gamtos pažinimo ir socialinių mokslų mokymosi sričių testų pasiekimais. Buvo keliama hipotezė, kad koreliacija su nacionaliniu reprezentaciniu testu turėtų būti panaši kaip ir valstijos atsiskaitomojo testo.

Antrasis tyrimo klausimas susijęs su žinių ir įgūdžių, kurie gali būti aktualūs taikant formuojamąjį vertinimo modelį, sudėtingumo analize. Dalykinės kalbos

rodikliai, pavyzdžiui, žodyno atitiktis, turinio suvokimas ir pan., nėra vieninteliai kintamieji, numatantys pasirinktų kriterijų pasiekimus, todėl kiti turinio rodikliai irgi turėtų būti vertinami, t. y. kai tikimasi, kad mokiniai išmoks ir įpras vartoti dalykinę kalbą, taip pat tikimasi, kad mokiniai skaitys ir supras mokomuosius tekstus ir su tuo susijusią medžiagą (*National Governors Association Center for Best Practices*, *Council of Chief State School Officers*, 2010), raštu apibendrins ir pritaikys tai, ką išmoko ir ką perskaitė.

Marcotte ir Hintze (2009) vertino skaitymo supratimo aiškinamuosius modelius, kurie sugretino skaitymo balsu sklandumą su teksto supratimo bendrųjų rezultatų vertinimu (pvz., metodas „Labirintas“), atpasakojimo sklandumu, teisingų ar klaidingų teiginių žymėjimo metodu ir rašytiniu atpasakojimu. Visų rezultatų vertinimo prognostinės koreliacijos su kriterijumi buvo vidutiniškai stiprios, varijuojančios nuo ,46 iki ,67. Daugybinės regresijos analizė parodė, kad keturių matavimų rinkinys (t. y. skaitymo balsu sklandumas, mokymo metodas „Labirintas“, teisingų ar klaidingų teiginių žymėjimo metodas ir žodinis atpasakojimas) sudarė apie 57 proc. kintamumo, vertinant skaitymo gebėjimą apskritai. Be to, visi „Labirinto“ mokymo metodo, teiginių žymėjimo metodo ir rašytinio atpasakojimo matavimai reikšmingai prisidėjo prie aiškinamojo modelio, atlikus skaitymo balsu sklandumo patikrą, tobulinimo. Šis tyrimas atskleidė, kad atlikti keturių kintamųjų matavimai yra geresnis prognozavimo mechanizmas nei pavienis atskirų kintamųjų matavimas.

Šiame tyrime trys matavimai, orientuoti į dalykinės kalbos vartojimą ir jos supratimą, buvo vertinami siekiant nustatyti, ar būtų didesnis skirtumas įsisavinant dalyko turinį daugeriopo matavimo atveju nei vertinant turinio supratimą. Potencialūs bendrų rezultatų matavimo įrankiai – teisingų ar klaidingų teiginių žymėjimo metodas ir rašytinis atpasakojimas – buvo pasirinkti vietoj skaitymo balsu sklandumo ir „Labirinto“ metodo taikymo. Ankstesni tyrimai parodė, kad dalykinės kalbos vertinimas stipriau koreliavo su kriterijumi, kai buvo lyginamas aprašomuoju ar tiesioginiu būdu (Espin ir Foegen, 1996; Mooney, McCarter, Schraven ir kt., 2013), pasirenkant jį kaip baigiamąją užduotį. Buvo keliama hipotezė, kad daugelio elementų vertinimo sistema yra geresnė nei vieno matavimo sistema, iš anksto numatant tyrimų rezultatą.

## Tyrimo dalyviai

Institucinio recenzavimo tarybos patvirtinto tyrimo dalyviai buvo vienos valstybinės Pietų Luizianos mokyklos, kurią lanko vaikai nuo ikimokyklinio amžiaus iki penktos klasės, penktos klasės mokiniai (N = 51). Šių vaikų tėvai sutiko,

kad vaikai dalyvautų tyrime. Tyrimo dalyvių amžius – 11,1 metų (standartinis nuokrypis SD = ,5) testavimo metu. Tyrime dalyvavo 68,6 proc. mergaičių (n = 35), 66,7 proc. baltaodžių (n = 34) ir 68,6 proc. gaunančiųjų nemokamus pietus (n = 35). 96 proc. tyrimo dalyvių (n = 49 abiem atvejais) pasiekė pagrindinį lygį pagal valstijos lygio atsiskaitomuosius testus (toks lygis traktuojamas kaip teigiamas, jei gamtos pažinimo ir socialinių mokslų testo rezultatai suskirstyti kategorijomis, pvz., anglų kalbos ir matematikos testai Luizianoje). Visoje valstijoje 43 proc. ir 47 proc. penktos klasės mokinių pasiekė pagrindinį gamtos pažinimo ir socialinių mokslų valstijos testo lygį.

## Matavimai

Šiame tyrime palyginti penki matavimai. Tam tikslui buvo naudojami gamtos pažinimo ir socialinių mokslų mokymosi turinio internetiniai sutrumpinti Stanfordo pasiekimų dešimtosios laidos testai (SAT-10; *Pearson Education*, n. d.) ir integruotoji Luizianos ugdymo vertinimo programa (iLEAP; LDE, n. d.). Numatomi kintamieji buvo dalyko turinio supratimo vertinimas, teisingų ar klaidingų teiginių žymėjimo metodas ir rašytinis atpasakojimas.

**SAT-10.** Sutrumpinta internetinė testo SAT-10 forma yra standartizuotas norminis pasiekimų testas, matuojantis mokinių nuo darželio iki 12 klasės skaitymo, matematikos, taisyklingos rašybos, kalbos, klausymosi, gamtos pažinimo ir socialinių mokslų mokymosi rodiklius. Šiame tyrime buvo analizuojami tik klasės lygmens gamtos pažinimo ir socialinių mokslų mokymosi testai. Testų sudarytojai apibūdino testus kaip atspindinčius esamą praktiką ir prilygino juos nacionaliniams ir valstijos standartams. Gamtos pažinimo testas vertino gyvybės, fizikos ir žemės mokslų žinias bei gamtos pažinimą kaip tyrinėjimą. Socialinių mokslų testas vertino istorijos, geografijos, politikos ir ekonomikos mokslų žinias. Kiekvienas sutrumpintas testas sudarytas iš 30 klausimų ir iš kelių pasirenkamųjų atsakymų, o laikas, skirtas jam atlikti, buvo neribojamas. Testo rezultatais naudojamasi šiame tyrime. Pamatuotas balas prilyginamas kiekvienam dalyko testui. Tokiu būdu galima stebėti kiekvienos klasės pasiekimų lygį (*Pearson Education*, n. d.). 2013 m. pavasario testavimo laikotarpiu testo imties vidutinis gamtos pažinimo balas buvo 658 (diapazonas nuo 602 iki 726), t. y. visos grupės duomenys atitiko 63 procentilį, lyginant su visos šalies duomenimis. Vidutinis socialinių mokslų mokymosi balas buvo 655 (diapazonas nuo 597 iki 713), tai atitiko 59 procentilį. Mokslininkai (Carney, Morse, n. d.), atlikę tyrimą apie mokinių, besimokančių valstybinėse mokyklose, pasiekimus, pateikė gautą tyrimo medžiagą naudodamiesi SAT-10 testu kaip visuma. Carney ir Morse apibūdino papildomus patikimumo ir turinio validumo įrodymus SAT-

10 atveju. Šiame tyrime kriterijaus validumas buvo įrodytas vidutiniškai stipria koreliacija tarp SAT-10 ir iLEAP turinio testų su ,64 ir ,69 linijinėmis sąsajomis gamtos pažinimo ir socialinių mokslų srityse (abiejų $p < ,01$).

**iLEAP penktos klasės kriterinis testas.** iLEAP tikslas yra matavimas, taikomas siekiant Luizianos mokymosi standartų anglų kalbos ir (ar) kalbos meno, matematikos, gamtos pažinimo ir socialinių mokslų srityse (LDE, n. d.) pritaikymo visiems trečios, penktos, šeštos, septintos ir devintos klasių mokiniams. Gamtos pažinimo ir socialinių mokslų testus sudarė klausimai ir keli pateikti atsakymų variantai. Klausimams atsakyti skirtas laikas nebuvo ribojamas. Tyrimai buvo atliekami skirtingomis dienomis. Penktos klasės gamtos pažinimo mokymosi turinys apėmė fizikos, gyvybės, žemės, kosmoso ir aplinkosaugos žinias; testo klausimai buvo formuluojami iš visų šių sričių. Socialinių mokslų mokymosi turinys apėmė geografiją, civilinę teisę, ekonomiką ir istoriją; testo klausimai buvo formuluojami iš geografijos ir istorijos sričių (LDE, n. d., a). Pasiekimų lygio aprašai buvo suskirstyti į nepatenkinamo, pagrindinio, gerai įvaldyto ir pažangaus lygio aprašus. iLEAP penktos klasės testų techninio adekvatumo duomenys buvo prieinami per LDE tinklalapį. Cronbacho alfa lygiai ,85 gamtos pažinimui ir ,82 socialiniams mokslams buvo nurodyti kaip 2010 m. testo vidinio vientisumo (angl. *consistency*) įrodymas (LDE, n. d., b). Pateikti duomenys buvo apibūdinti kaip turinio validumas, kuris nebuvo aprašytas (LDE, n. d., b).

**Dalyko turinio suvokimas.** Turinio suvokimo rezultatų matavimas išsirutuliojo iš procedūrų, kurias anksčiau apibrėžė Espin, Busch, Shin ir Kruschwitz (2001). Kiekviename tyrime sąvokos buvo atsitiktine tvarka atrinktos iš viso dalyko turinio sąvokų sąrašo, pateikto vadovėlyje, ir peržiūrėtos mokytojų grupės (atrinkti mokytojai buvo vertinami kaip sumanūs ir geri savo srities specialistai). Sąvokų sąrašas buvo sudarytas remiantis mokymosi programa. Kiekvienas tyrimas apėmė skirtingos mokymosi programos dalies sąvokas. Sąvokos iš kiekvienos mokymo dalies buvo parinktos paskaičiavus metų mokymo programos proporciją, skirtą kiekvienai daliai pagal valstijos mokymosi spartos vadovą, tada dauginama iš kiekvieno tyrimo klausimų skaičiaus. Dvidešimt sąvokų ir su jomis susijusių apibrėžimų buvo pateikti internetinėje sistemoje („Moodle“, n. d.), kartu pateikti ir keli atsakymų variantai į kiekvieną klausimą. Kintanti formos patikimumo koreliacija 20-čiai lygiagrečių tyrimų vidutiniškai buvo ,55 (varijavo nuo ,21 iki ,73; standartinis nuokrypis SD = ,09) (Mooney, McCarter, Russo ir kt., 2013). Kriterijaus validumo rezultatai buvo pateikti anksčiau.

**Teisingų ar klaidingų teiginių žymėjimo metodas** sukurtas siekiant įvertinti skaitomo teksto supratimą (Royer ir kt., 1979). Šio tyrimo eigą sudarė

teksto skaitymas, po to tyrimo dalyviams buvo pateikti sakiniai, kuriuos jie turėjo skaityti jau nematydami teksto. Pateikti įvairių tipų sakiniai: (a) iš skaityto teksto paimti sakiniai; (b) perfrazuoti ar panašios prasmės sakiniai su pakeistais žodžiais; (c) pakeistos prasmės ar panašūs sakiniai su nedideliais žodžių pakeitimais, keičiančiais sakinių prasmę, (d) atitraukiantys dėmesį sakiniai ar panašūs pagal teksto turinį sakiniai, kurie nuo skaitytos teksto ištraukos skiriasi ir prasme, ir formuluote. Šiame tyrime teksto turinys parengtas pagal gamtos pažinimo ir socialinių mokslų sričių tekstus ir pagal klasės lygį. Tekstai padalyti pagal gamtos pažinimo ir socialinių mokslų sričių turinį ir pristatyti pakaitomis (16 sakinių prie kiekvieno teksto). Tyrimo dalyvis, perskaitęs tekstą, turėdavo žymėti sakinius „taip“ (t. y. sakinio prasmė atitinka teksto prasmę) arba „ne“ (t. y. prasmė skiriasi). Kriterijaus validumo koreliacijos su standartizuoto testo matavimais, įskaitant SAT, varijavo nuo ,50 iki ,73; keturių ištraukų tyrimo patikimumas varijavo nuo ,70 iki ,80 (Royer, 2004).

**Rašytinis atpasakojimas.** Rašytinio atpasakojimo tyrimas buvo identiškas tam, kurį taikė Marcotte ir Hintze (2009). Mokinių buvo prašoma per 5 minutes tyliai perskaityti 750 žodžių tekstą. Tada mokiniai turėjo užrašyti viską, ką jie atsimena. Balai už rašytinį atpasakojimą buvo skiriami už pakartotus turinio žodžius. Pasak Marcotte ir Hintze, turinio žodžiai buvo apibrėžti kaip „atskiri tikriniai ir bendriniai daiktavardžiai, veiksmažodžiai, būdvardžiai ir prieveiksmiai, esantys tekste, arba sinonimai tiems, kurie paminėti tekste“ (p. 322). Turinio žodžių sąrašas buvo parengtas remiantis skaitomu tekstu. Kriterijaus validumo koreliacija rašytinio atpasakojimo standartizuoto pasiekimo testo atveju buvo ,57 (Marcotte ir Hintze). Rašytinio atpasakojimo tyrimas parodė atitinkamai ,56 ir ,59 koreliaciją su skaitymo balsu sklandumo ir „Labirinto“ metodo užduotimis.

## Procedūros

Testavimas vyko gamtos pažinimo pamokos metu 2013 m. gegužės viduryje, baigiantis mokslo metams ir praėjus šešioms savaitėms po valstijos atsiskaitomojo testo laikymo. SAT-10 internetinio testavimo metu kilusios techninės problemos lėmė tai, kad šis testas visų dalyvių buvo atliktas paskutinis. Testą, prisijungę prie saugių interneto svetainių, mokiniai atliko stebimi pirmojo straipsnio autoriaus. Teisingų ar klaidingų teiginių žymėjimo metodo ir rašytinio atpasakojimo atvejais testavimui vadovavo pirmasis ir ketvirtasis straipsnio autoriai, pastarasis yra gamtos pažinimo dalyko mokytojas.

### Rezultatų balų suderinimas

Teisingų ar klaidingų teiginių žymėjimas ir rašytinis atpasakojimas buvo įvertinti balais atskirai pirmojo ir antrojo straipsnio autorių. Balai registruojami duomenų bazėje. Pirminė duomenų analizė apėmė visų dalyvių galutinių balų suderinimo proporcijų skaičiavimą. Suderinimas nustatomas tada, kai mokinių individualių tyrimų galutiniai balai yra vienodi. Suderinimo proporcijos buvo 100 proc. teisingų ar klaidingų teiginių žymėjimo atveju ir 85,7 proc. rašytinio atpasakojimo atveju. Pirmojo autoriaus pirmieji balai buvo naudojami tyrimo analizei. Jokie rezultatų suderinimo veiksmai nebuvo naudojami vertinant turinio supratimą. SAT-10 testo balai buvo pateikti testo sudarytojų, o iLEAP testo balus pateikė mokytojas.

### Analizė

Analizė apėmė kriterijaus validumo klausimus. Kriterijaus validumas kiekvienam iš trijų numatomų kintamųjų buvo vertinamas taikant koreliaciją ir 95 proc. pasikliautinumo intervalą (CI) tarp kiekvieno tyrimo ir aktualaus kriterijaus matavimo. Visi kintamieji buvo įvertinti ir laikomi normaliai pasiskirstę pritaikius Šapiro ir Vilko (Shapiro ir Wilk) testą. Testo validumas buvo įvertintas naudojant nuoseklią daugybinę linijinę regresiją ir panašumų analizę. Gamtos pažinimo ir socialinių mokslų dalykų turinys buvo analizuojamas pasirinkus atskirą nuoseklios (ar hierarchinės) daugybinės regresijos analizę. Nuosekli regresija buvo taikoma siekiant nustatyti, ar papildomų kintamųjų teikiama informacija pagerino kriterijaus spėjimo vertinimą po to, kai buvo statistiškai eliminuoti anksčiau suregistruoti kintamieji (Tabachnick ir Fidell, 2013). Siekiant nustatyti, kokius pokyčius lėmė supratimo matavimas, turinio suvokimo vertinimo rodikliai buvo suregistruojami pirmiausia, po to – teisingų ar klaidingų teiginių žymėjimo ir rašytinio atpasakojimo rodikliai. Galiausiai panašumų analizė buvo atlikta siekiant nustatyti kriterijaus kintamųjų įvairovę (t. y. SAT-10 testai gamtos pažinimo ir socialinių mokslų srityse), priklausomai nuo numanomų kintamųjų. Panašumų formulės „Excel“ programos formatu buvo paimtos iš Warne (2011).

## Rezultatai

### Kriterijaus validumas

1 lentelė demonstruoja valstijos ir šalies dalyko turinio supratimo testų imčių vidurkius, skirstinius, 95 proc. pasikliautinumo intervalus ir kiekvieną iš trijų rodiklių. Normalumo patikrinimas parodė, kad visi gamtos pažinimo daly-

ko balai, išskyrus turinio suvokimą, buvo normaliai pasiskirstę. 2 lentelė rodo visų vertinimų koreliacijas su 95 proc. pasikliautinumo intervalais. Koreliacijos, 95 proc. pasikliautinumo intervalai tarp rodiklių ir kriterijaus kintamieji buvo pasiskirstę nuo žemo (t. y. < ,3) iki vidutinio (t. y. nuo ,3 iki ,7) diapazono. Turinio suvokimas labiausiai koreliavo su abiem turinio testais, su teisingo ar klaidingo teiginio žymėjimu ir rašytiniu atpasakojimu. Socialinių mokslų mokymosi koreliacija buvo didesnė nei gamtos pažinimo mokymosi koreliacija. Turinio suvokimo ir teisingų ar klaidingų teiginių žymėjimo metodų koreliacijos ir kriterijų matavimai reikšmingai skyrėsi nuo nulio ($p < ,01$). Teisingų ar klaidingų teiginių žymėjimas parodė stipresnę koreliaciją su socialinių mokslų mokymosi vertinimo testais nei gamtos pažinimo testais, o rašytinis atpasakojimas reikšmingai koreliavo tik su valstijos socialinių mokslų mokymosi vertinimo testu.

1 lentelė

**Valstijos atsiskaitomųjų testų, standartizuotų testų ir numatomo rodiklio testų vidurkiai penktoje klasėje**

| | N | Diapazonas | Vidurkis | Standartinis nuokrypis (SD) | 95 proc. pasikliautinumo intervalas (CI) | Asimetrija | Ekscesas |
|---|---|---|---|---|---|---|---|
| iLEAP, Gamtos pažinimas | 51 | 276–454 | 360,6 | 36,8 | 350, 371 | ,33 | ,62 |
| iLEAP, Socialiniai mokslai | 51 | 281–424 | 347,3 | 31,7 | 338, 356 | ,29 | -,01 |
| SAT-10, Gamtos pažinimas | 51 | 602–726 | 658,5 | 25,6 | 651, 666 | ,15 | ,29 |
| SAT-10, Socialiniai mokslai | 46 | 586–713 | 654,0 | 27,7 | 646, 662 | ,06 | ,17 |
| TS, Gamtos pažinimas | 49 | 8–20 | 15,73 | 3,05 | 14,86, 16,61 | -0,80 | 0,02 |
| TS, Socialiniai mokslai | 50 | 3–18 | 11,34 | 4,34 | 10,08, 12,60 | -0,08 | -1,08 |
| TKP | 50 | 29–59 | 47,36 | 6,54 | 45,50, 49,22 | -0,32 | 0,34 |
| RA | 49 | 11–42 | 26,92 | 7,86 | 24,66, 29,18 | 0,39 | -0,78 |

*Pastaba.* TS – turinio suvokimas; TKP – teisingų ar klaidingų teiginių žymėjimo metodas; RA – rašytinis atpasakojimas; iLEAP – integruota Luizianos ugdymo vertinimo programa; SAT-10 – Stanfordo pasiekimų testas, dešimtoji laida, sutrumpinta forma.

2 lentelė

**Koreliacijos ir 95 proc. pasikliautinumo intervalai tarp visų numatymų ir kriterijaus matavimų**

| Testas | iLEAP, Socialiniai mokslai | SAT, Gamtos pažinimas | SAT, Socialiniai mokslai | TS, Gamtos pažinimas | TS, Socialiniai mokslai | TKP | RA |
|---|---|---|---|---|---|---|---|
| iLEAP, Gamtos pažinimas | ,66** [,47, ,79] | ,64** [,44, ,78] | ,57** [,35, ,73] | ,51** [,27, ,69] | ,47** [,22, ,66] | ,46** [,21, ,65] | ,24 [-,03, ,48] |
| iLEAP, Socialiniai mokslai | | ,49** [,25, ,68] | ,69** [,54, ,81] | ,61** [,40, ,76] | ,66** [,47, ,79] | .49** [,25, ,68] | ,29* [,02, ,52] |
| SAT, Gamtos pažinimas | | | ,51** [,27, ,69] | ,55** [,32, ,72] | ,43** [,18, ,63] | ,49** [,25, ,68] | ,16 [-,12, ,42] |
| SAT, Socialiniai mokslai | | | | ,65** [,26, ,70] | ,63** [,42, ,77] | ,59** [,36, ,75] | ,28 [-,01, ,53] |
| TS, Gamtos pažinimas | | | | | ,60** [,38, ,75] | ,40** [,13, ,61] | ,26 [-,02, ,50] |
| TS, Socialiniai mokslai | | | | | | ,41** [,15, ,62] | ,25 [-,03, ,49] |
| TKP | | | | | | | ,22 [-,06, ,54] |

*Pastaba.* ** Koreliacija yra reikšminga ,01 lygiu; * koreliacija yra reikšminga ,05 lygiu. TS – turinio suvokimas; TKP – teisingų ar klaidingų teiginių žymėjimo metodas; RA – rašytinis atpasakojimas; iLEAP – integruota Luizianos ugdymo vertinimo programa; SAT – Stanfordo pasiekimų testas, dešimtoji laida, sutrumpinta forma.

## Validumas

Remiantis koreliacijų analizės rezultatais, numatomos rodiklių reikšmės buvo surašytos į nuoseklų regresijos modelį: pirmiausia turinio suvokimo vertinimas, po to teisingų ar klaidingų teiginių žymėjimas ir rašytinis atpasakojimas. Duomenys buvo analizuoti, siekiant nustatyti reikšmingus taškus ir daugialypį bendrą linijiškumą. Du stebėjimai parodė Cooko D vertes, didesnes už 1, regresijos analizė buvo atlikta su jomis ir be jų – kadangi rezultatai žymiai nepakito, stebėjimų informacija buvo išsaugota duomenų rinkinyje. Formalūs daugialy-

pio linijiškumo testai nieko neatskleidė. Likutinės vertės parodė pakitimų homogeniškumą ir linijinius ryšius tarp visų prediktorių ir kriterijaus kintamųjų. 3 ir 4 lentelės rodo regresinės analizės rezultatus trijų kintamųjų ir kriterijaus SAT-10 testo atvejais. Rezultatai parodė, kad daugiausia pakitimų yra susiję su turinio supratimo vertinimu ir teisingų ar klaidingų teiginių žymėjimu. Rašytinis atpasakojimas nebuvo reikšmingas rodiklis nei gamtos pažinimo, nei socialinių mokslų mokymosi vertinimo atveju, koreguotam $R^2$ sumažėjus abiejuose SAT-10 testų rezultatuose.

3 lentelė

**Nuoseklioji regresija SAT-10 (gamtos pažinimas) su numatymo rodikliais, suregistruotais pagal pateikimo eiliškumą**

| | | | | | 1 modelis | | 2 modelis | | 3 modelis | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Numatymas** | **R** | **$R^2$** | **Koreg. $R^2$** | **$R^2$ pokytis** | **β** | **p vertė** | **β** | **p vertė** | **β** | **p vertė** |
| TS | 0,54 | 0,30 | 0,28 | 0,30 | 4,83 | ,000 | 3,75 | ,002 | 3,79 | ,002 |
| TS, TKP | 0,61 | 0,37 | 0,35 | 0,08 | | | 1,21 | ,023 | 1,22 | ,024 |
| TS, TKP, RA | 0,61 | 0,37 | 0,33 | 0,00 | | | | | -0,08 | ,855 |

*Pastaba.* SAT-10 – Stanfordo pasiekimų testas, dešimtoji laida, sutrumpinta forma; TS – turinio suvokimas; TKP – teisingų ar klaidingų teiginių žymėjimo metodas; RA – rašytinis atpasakojimas.

4 lentelė

**Nuoseklioji regresija, SAT-10 (socialiniai mokslai) su numatymo rodikliais, suregistruotais pagal pateikimo eiliškumą**

| | | | | | 1 modelis | | 2 modelis | | 3 modelis | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Numatymas** | **R** | **$R^2$** | **Koreg. $R^2$** | **$R^2$ pokytis** | **β** | **p vertė** | **β** | **p vertė** | **β** | **p vertė** |
| TS | 0,62 | 0,38 | 0,36 | 0,38 | 3,64 | ,000 | 2,73 | ,000 | 2,65 | ,000 |
| TS, TKP | 0,72 | 0,52 | 0,50 | 0,14 | | | 1,76 | ,001 | 1,72 | ,001 |
| TS, TKP, RA | 0,73 | 0,53 | 0,49 | 0,01 | | | | | 0,23 | ,452 |

*Pastaba.* SAT-10 – Stanfordo pasiekimų testas, dešimtoji laida, sutrumpinta forma; TS – turinio suvokimas; TKP – teisingų ar klaidingų teiginių žymėjimo metodas; RA – rašytinis atpasakojimas.

5 lentelė rodo panašumų analizės rezultatus. Gamtos pažinimo atveju regresijai poveikį labiausiai darė turinio supratimas (14,8 proc.) ir teiginių žymėjimo

metodas (7,3 proc.). Tyrimas neatskleidė pamatuojamo pokyčio, kuris būtų priskirtas rašytiniam atpasakojimui. Turinio supratimas ir teisingų ar klaidingų teiginių žymėjimas sudarė trečdalį pasikeitimų kriterijaus kintamajame (14,8 + 7,3 + 12,8 = 34,9 proc.). Analizuojamas kartu su turinio supratimo vertinimu, rašytinis atpasakojimas atliko turinio suvokimo slopinimo poveikį, kurį parodė neigiamas koeficientas. 5 lentelė rodo ir tai, kad SAT-10 socialinių mokslų srities rezultatų regresijos efektą labiausiai lėmė turinio suvokimo (17,3 proc.) ir teiginių žymėjimo metodas (11,3 proc.). Rašytinio atpasakojimo, remiantis tyrimo rezultatais, pokytis buvo slopinamasis (–1,6 proc. nuo visos apimties). Analizuojant kartu turinio suvokimą ir teiginių žymėjimo metodą, pakitimai sudarė 44,9 proc. visų SAT-10 testo pakitimų socialinių mokslų srityje (17,3 + 11.3 + 16,3).

5 lentelė

**Dažnumo analizė SAT-10 (gamtos pažinimas ir socialiniai mokslai)**

| **SAT, Gamtos pažinimas** | **TS** | **TKP** | **RA** | **$R^2$ atskyrimas** | **SAT, Socialiniai mokslai** | **TS** | **TKP** | **RA** | **$R^2$ atskyrimas** |
|---|---|---|---|---|---|---|---|---|---|
| $U_1$ | 14,8 % | -- | -- | 14,8 % | $U_1$ | 17,3 % | -- | -- | 17,3 % |
| $U_2$ | -- | 7.3 % | -- | 7,3 % | $U_2$ | -- | 11,3 % | -- | 11,3 % |
| $U_3$ | -- | -- | 0,0 % | 0,0 % | $U_3$ | -- | -- | -1,6 % | -1,6 % |
| $D_{12}$ | 12,8 % | 12,8 % | -- | 12,8 % | $D_{12}$ | 16,3 % | 16,3 % | -- | 16,3 % |
| $D_{13}$ | -1,0 % | -- | -1,0 % | -1,0 % | $D_{13}$ | 1,7 % | -- | 1,7 % | 1,7 % |
| $D_{23}$ | -- | 0,1 % | 0,1 % | 0,1 % | $D_{23}$ | -- | 3,1 % | 3,1 % | 3,1 % |
| $D_{123}$ | 3,4 % | 3,4 % | 3,4 % | 3,4 % | $D_{123}$ | 4,4 % | 4,4 % | 4,4 % | 4,4 % |
| Sum = $r^2$ | 30,0 % | 23,6 % | 2,5 % | -- | | 39,7 % | 35,1 % | 7,6 % | -- |
| Sum = $R^2$ | | | | 37,4 % | | | | | 52,5 % |
| Unikalūs | 14,8 % | 7,3 % | 0,0 % | | Unikalūs | 17,3 % | 11,3 % | -1,6 % | |
| Panašūs | 15,2 % | 16,3 % | 2,5 % | | Panašūs | 22,4 % | 23,8 % | 9,2 % | |
| Iš viso | 30,0 % | 23,6 % | 2,5 % | | Iš viso | 39,7 % | 35,1 % | 7,6 % | |

*Pastaba.* SAT-10 – Stanfordo pasiekimų testas, dešimtoji laida, sutrumpinta forma; TS – turinio suvokimas; TKP – teisingų ar klaidingų teiginių žymėjimo metodas; RA – rašytinis atpasakojimas; U – unikalūs; D – dažni; 1 – TS; 2 – TKP; 3 – RA.

## Diskusija

Validavimas yra būtina sąlyga rengiant testus (*American Educational Research Association*, *American Psychological Association*, *National Council on Measurement in Education*, 1999). Validumas labai svarbus specialiesiems peda-

gogams, atsakingiems už ugdymo programų individualizavimą ir tinkamą jų pritaikymą praktikoje. Šis tyrimas pateikė teksto turinio supratimo kriterijaus validumo įrodymus, taip pat ir pirminius įrodymus, susijusius su instrumentų paketu, taikomu atliekant formuojamąjį valstybinių mokyklų mokinių socialinių mokslų ir gamtos pažinimo žinių vertinimą.

Turinio suvokimo vertinimo balai vidutiniškai koreliavo (t. y. nuo ,3 iki ,7) gamtos pažinimo ir socialinių mokslų atvejais su balais šalyje pripažinto SAT-10 testo; to ir buvo tikimasi. Gamtos pažinimo srityje koreliacija ,55 (95 proc. pasikliautinumo indeksas; ,32, ,72) apėmė intervalą nuo ,36 iki ,55, esant linijiniam ryšiui su valstijos testu, anksčiau aprašytu Mooney, McCarter, Russo ir kt. (2013). Vidutinė koreliacija tarp teksto turinio supratimo ir valstijos lygio atsiskaitomojo testo (t. y. ,51) vėl pasikartojo (žr. 2 lentelę). Socialinių mokslų srityje koreliacija su SAT-10 testu buvo ,63 (95 proc. pasikliautinumo indeksas; ,42, ,77), t. y. buvo aukštesnė negu gamtos pažinimo koreliacija ir panaši į ,68 linijinį ryšį su valstijos atsiskaitomuoju testu, aptartu Mooney ir kt. (2014), ir šiame tyrime siekė ,66 (žr. 2 lentelę). Turinio suvokimo vertinimo sąsajų, kurios buvo didesnės socialinių mokslų srityje nei gamtos pažinimo, modelis buvo taikomas šalyje ir valstijoje atliekamuose vertinimuose. Tolesni turinio suvokimo rezultatų validumo įrodymai buvo aiškinamieji. Regresijos modeliuose pakitimų kiekio identifikavimas buvo atliktas dėl numatytų kintamųjų; turinio suvokimas buvo stipriausias rodiklis tiek gamtos pažinimo, tiek socialinių mokslų SAT-10 testų atvejais. Be to, panašumų analizė parodė, kad nuo 15 proc. iki 17 proc. pakitimų modeliuose buvo unikalūs, paaiškinami turinio suvokimu.

Šio tyrimo rezultatai identifikavo teisingų ar klaidingų teiginių žymėjimo metodą kaip potencialų mokinių veiklos rodiklį mokomųjų dalykų turinio prasme, nuo 7 proc. iki 11 proc. pakitimų buvo paaiškinta vien teiginių žymėjimo metodu, ir maždaug 35–45 proc. pakitimų buvo paaiškinti kartu sudėjus turinio suvokimo ir teisingų ar klaidingų teiginių žymėjimo metodus. Rašytinis atpasakojimas neprisidėjo prie regresijos poveikio ir turėjo mažai panašių pakitimų, lyginant su kriterijaus kintamuoju. Abu anksčiau buvo įvertinti kaip potencialūs skaitymo supratimo matavimai, aprašyti Marcotte ir Hintze (2009). Teisingų ar klaidingų teiginių žymėjimo metodas anksčiau buvo traktuojamas kaip turintis vidutinę koreliaciją su kriterijais ir aiškinamaisiais gebėjimais; dar anksčiau buvo nustatyta, kad šis metodas vidutiniškai koreliuoja su skaitymo supratimo modelių kriterijais ir aiškinamaisiais gebėjimais (Marcotte ir Hintze), vidutiniškai koreliuoja su gamtos pažinimo ir socialinių mokslų turinio mokymosi testų balais. SAT-10 atveju – su linijiniais ryšiais ,49 (95 proc. pasikliautinumo indeksas; ,25 ,67) gamtos pažinimo ir ,59 (95 proc. pasikliautinumo indeksas; ,37, ,75) socialinių mokslų mokymosi srityse. Rezultatai parodė, kad dalyko turinio

žodynas ir skaitomo teksto supratimas yra reikšmingi rodikliai, galintys paveikti mokinių gamtos ir socialinių mokslų sričių pasiekimus, todėl gali būti vertinami kaip numatymo rodikliai (prediktoriai) tiek valstijos, tiek ir šalies pasiekimų testuose. Rašytinio atpasakojimo atveju šiame tyrime nė viena iš koreliacijų su SAT-10 testu nebuvo vidutinio dydžio ar statistiškai reikšminga. Gauti rezultatai skiriasi nuo aprašytų Marcotte ir Hintze (2009), nustačiusių, kad rašytinis atpasakojimas buvo reikšmingas rodiklis skaitymo supratimo modeliuose. Skirtingus rezultatus rašytinio atpasakojimo atveju galėjo lemti tai, kad rašytinis atpasakojimas buvo vertinamas pagal reikšminius žodžius, kuriuos mokinys iš perskaityto teksto prisimena ir juos užrašo. Gali būti, kad, mokydamiesi gamtos pažinimo, mokiniai yra šiek tiek mažiau susipažinę su dalyko žodynu, taigi dėl šios priežasties jiems sunkiau sekasi prisiminti ir užrašyti žodžius.

## Tyrimo apribojimai

Atliekant šį tyrimą ir įgyvendinant pirminį tyrimo planą dėl techninių sunkumų kilo tam tikrų apribojimų. Pirma, kadangi testavimo tvarka buvo pakeista, visų mokinių testavimas SAT-10 testu vyko paskiausiai, o tai galėjo padaryti poveikį tyrimo rezultatams. Antra, dalyvių imtis buvo patogioji, gavus leidimą, apimanti aukštus mokymosi rezultatus demonstruojančius mokinius, ir tai galėjo būti nereprezentatyvu. Palyginus imtis šalies mastu, dalyvių grupės vidurkis šioje imtyje buvo 63 ir 59 procentiliai atitinkamai SAT-10 testo gamtos pažinimo ir socialinių mokslų srityse. Trečia, imtį sudarė vienos klasės mokiniai, todėl reikia būti labai atidiems apibendrinant tyrimo rezultatus.

## Implikacijos

Turint omenyje šiuos apribojimus ir aprašytus tyrimo rezultatus, verta pastebėti, kad įrodymai pagrindžia svarstymus apie dalykinės kalbos rodiklius kaip perspektyvius bendrų rezultatų matavimo įrankius, fiksuojant mokinių veiklą ir pažangą mokantis gamtos pažinimo ir socialinių mokslų. Koncentruojantis į pirmąjį (iš trijų) etapą ir balų vertinimą (Fuchs, 2004), gauti rezultatai padeda kurti dalyko turinio vertinimo schemą.

Pirmoji implikacija susijusi su dalyko žodyno matavimo panaudojimu. Lyginamasis tyrimas atliekant bendrų rezultatų matavimą turinio mokymosi atveju palankiai vertina tokius instrumentus kaip turinio suvokimas, žodyno atitiktis ir pagrindinis žodynas. Espin ir Foegen (1996) atliko trijų rodiklių (t. y. žodyno atitikties, „Labirinto“ mokymo metodo ir skaitymo balsu sklandumo) palyginimą ir padarė išvadą, kad stipriausia koreliacija su kriterijumi buvo žodyno ati-

tikties atveju. Mooney, McCarter, Schraven ir kt. (2013) atliko tiesioginius palyginimus tarp žodyno atitikties, „Labirinto“ mokymo metodo ir rašymo rezultatų matavimo koreliacijų ir nustatė statistiškai reikšmingus koreliacijos skirtumus su valstijos atsiskaitomuoju testu, kuriam gerai tiko žodyno atitikties matavimas. Šiame tyrime daugybinės regresijos analizė pateikė įrodymų, kad dalykinės kalbos tyrimas geriausiai paaiškina analizuotų modelių grupes. Toks pat modelis buvo pakartotas šiame tyrime vertinant turinio suvokimą.

Be to, bendrųjų rezultatų matavimas, dalykinę kalbą traktuojant kaip rodiklį, yra naudingas mokantis ir gamtos pažinimo, ir socialinių mokslų. Šios srities pirminis tyrimas buvo orientuotas į socialinių mokslų mokymąsi. Esant linijiniams ryšiams su kriterijaus matavimu visose aukštesnėse pradinio ir vidurinio ugdymo klasėse, buvo nustatyta, kad koreliacijos varijavo nuo vidutinių iki stiprių (pvz., Espin ir kt., 2001). Nepaisant to, naujausi tyrimai, įskaitant ir šį, praplėtė linijinių sąsajų vertinimą iki gamtos pažinimo mokymosi visose pradinio ir vidurinio ugdymo klasėse (pvz., Espin ir kt., 2013): gauti vidutiniai ir stiprūs statistiškai reikšmingi koreliaciniai ryšiai, varijuojantys nuo ,45 iki ,66. Tai, kad bendrųjų rezultatų matavimo indeksas, mokantis dalyko, gali būti taikomas dalykinėje sistemoje, patvirtina šio įrankio potencialą ir leidžia pripažinti jį kaip svarbų atsaką į intervenciją vertinimo schemose. Be to, logistinis įgyvendinamumas yra akivaizdus, nes testas gali būti administruojamas internetinėmis technologijomis. Tokie tyrimo rezultatai suteikia specialiesiems pedagogams tyrimu pagrįstą veikimo būdą, kai norima pamatuoti mokinių akademinius pasiekimus mokantis gamtos pažinimo ir socialinių mokslų.

Antroji implikacija susijusi su vertinimo technologijomis, taikomomis aukštesnėse pradinio ir vidurinio ugdymo klasėse. Bendrųjų rezultatų matavimo svarba per visą pradinio ir vidurinio ugdymo laikotarpį tik didėja. Bendrųjų rezultatų matavimo įrankiai ir vertinimo sistemos ir toliau yra tobulinami. Neseniai buvo aprašytas (Deno ir kt., 2009) mokyklos pradinių klasių skaitymo vertinimo programos, apėmusios bendrąjį žinių patikrinimą ir pažangos stebėseną, naudojant „Labirinto“ mokymo metodą ir skaitymo balsu sklandumo matavimą, įgyvendinimas. Buvo siekiama užfiksuoti pažangą įvairiose klasėse per tam tikrą laiką. Panašios pastangos išplėtoti ilgalaikes ir tikslias vertinimo technologijas, mokantis dalykų, pasiteisino.

Įvairialypiai matavimai yra pravartūs toliau tęsiant tyrimus. Mokantis gamtos pažinimo ir socialinių mokslų, turinio suvokimo vertinimas labiausiai paaiškino regresijos modelių atvejus. Teisingų ar klaidingų teiginių žymėjimo metodas, skaitymo ir turinio supratimo matavimas taip pat buvo reikšmingas rodiklis gamtos ir socialinių mokslų srityse. Bendrumų analizė (žr. 5 lentelę)

parodė, kad abu matavimai paaiškino gamtos pažinimo ir socialinių mokslų mokymosi sričių SAT-10 testo atlikimą. Teisingų ar klaidingų teiginių žymėjimo metodo tyrimas leido įvertinti skaitomo teksto supratimo modelius, kai kuriuos mokymo metodus: „Labirintą“, skaitymo balsu sklandumą ir rašytinį atpasakojimą (Marcotte ir Hintze, 2009). Mokinių mokymosi pokyčius galima tirti pasitelkiant daugybę matavimų, todėl būtina ištirti įvairių matavimų kombinacijų veiksmingumą, turint omenyje jų techninį pritaikymą, mokymo veiksmingumą, logistinį įgyvendinamumą (Deno ir Fuchs, 1987). Toks tyrimas formuotų ir galimai pateiktų informacijos apie įgyvendinamą tyrimą, atliekant vidurinių klasių mokinių skaitymo pažangos vertinimą (pvz., Barth ir kt., 2012; Tolar, Barth, Fletcher, Francis, Vaughn, 2014), ir galimai darytų poveikį tiek bendrajam, tiek ir specialiajam ugdymui. Specialieji pedagogai gali prisidėti prie atliekamų tyrimų savo darbe taikydami įvairias alternatyvias praktikas.

## Literatūra

Alexander, F. *Understanding Vocabulary*. Prieiga internete: http://www.scholastic.com/teachers/article/understanding-vocabulary

Barth, A. E., Stuebing, K. K., Fletcher, J. M., Cirino, P. T., Francis, D. J., & Vaughn, S. (2012). Reliability and Validity of the Median Score When Assessing the Oral Reading Fluency of Middle Grade Readers. *Reading Psychology*, *33*, 133–161.

Bloom, B. S. (1980). The New Direction in Educational Research: Alterable Variables. *Phi Delta Kappan*, *61*, 382–385.

Carney, R. N. Review of the Stanford Achievement Test, 10th ed. *Mental Measurements Yearbook*. Prieiga internete: http://buros.org/mental-measurements-yearbook

Deno, S. L. (1985). Curriculum-Based Measurement: The Emerging Alternative. *Exceptional Children*, *52*, 219–232.

Deno, S. L., & Fuchs, L. S. (1987). Developing Curriculum-Based Measurement Systems for Data-Based Special Education Problem Solving. *Focus on Exceptional Children*, *19* (8), 1–16.

Deno, S. L., Reschly, A. L., Lembke, E. S., Magnusson, D., Callender, S. A, Windram, H., et al.(2009). Developing a School-Wide Progress-Monitoring System. *Psychology in the Schools*, *46* (1), 46–55. doi: 10.1002/pits.20353.

Espin, C. A., Busch, T. W., Lembke, E. S., Hampton, D. D., Seo, K., & Zukowski, B. A. (2013). Curriculum-Based Measurement in Science Learning: Vocabu-

lary-Matching as an Indicator of Performance and Progress. *Assessment for Effective Intervention*, *38*, 203–213. doi: 10.1177/1534508413489724.

Espin, C. A., Busch, T. W., Shin, J., & Kruschwitz, R. (2001). Curriculum-Based Measurement in the Content Areas: Validity of Vocabulary Matching as an Indicator of Performance in Social Studies. *Learning Disabilities Research & Practice*, *16*, 142–151. Prieiga internete: http://dx.doi.org/10.1111/0938-8982.00015

Espin, C. A., & Deno, S. L. (1994–1995). Curriculum-Based Measures for Secondary Students: Utility and Task Specificity of Text-Based Reading and Vocabulary Measures for Predicting Performance on Content Area Tasks. *Diagnostique*, *20*, 121–142.

Espin, C. A., & Foegen, A. (1996). Validity of General Outcome Measures for Predicting Secondary Students' Performance on Content-Area Tasks. *Exceptional Children*, *62*, 497–514.

Espin, C. A., Shin, J., & Busch, T. W. (2005). Curriculum-Based Measurement in the Content Areas: Vocabulary Matching as an Indicator of Progress in Social Studies Learning. *Journal of Learning Disabilities*, *38*, 353–363. Prieiga internete: http://dx.doi.org/10.1177/00222194050380041301

Fuchs, L. S. (2004). The Past, Present, and Future of Curriculum-Based Measurement Research. *School Psychology Review*, *33*, 188–192.

Fuchs, L. S., & Deno, S. L. (1991). Paradigmatic Distinctions between Instructionally Relevant Measurement Models. *Exceptional Children*, *57*, 488–500.

Jenkins, J. R., & Fuchs, L. S. (2012). Curriculum-Based Measurement: The Paradigm, History, and Legacy. In C. A. Espin, K. L. McMaster, S. Rose, & M. M. Wayman (Eds.), *A Measure of Success: The Influence of Curriculum-Based Measurement on Education* (pp. 7–23). Minneapolis, MN: University of Minnesota Press.

Louisiana Department of Education. (LDE; a). *Integrated Louisiana Educational Assessment Program* (*i*LEAP). Prieiga internete: http://www.louisianaschools.net/lde/uploads/9725.pdf

Louisiana Department of Education. (LDE; b). *iLEAP 2010 Technical Summary.* Prieiga internete: http://www.louisianaschools.net/lde/uploads/18005.pdf

Marcotte, A. M., & Hintze, J. M. (2009). Incremental and Predictive Validity of Formative Assessment Methods of Reading Comprehension. *Journal of School Psychology*, *47*, 315–335. doi: 10.1015/j.jsp.2009.04.003.

Moodle. Prieiga internete: http://docs.moodle.org/23/en/About_Moodle

Mooney, P., McCarter, K. S., Russo, R. J., & Blackwood, D. L. (2013). Examining an Online Content General Outcome Measure Technical Features of the Static Score. *Assessment for Effective Intervention*, *38* (4), 249–260.

Mooney, P., McCarter, K. S., Russo, R. J., & Blackwood, D. L. (2014). The Structure of an Online Assessment of Science and Social Studies Content: Testing Optional Formats of a General Outcome Measure. *Social Welfare Interdisciplinary Approach*, *4* (1), 81–93.

Mooney, P., McCarter, K. S., Schraven, J., & Callicoatte, S. (2013). Additional Performance and Progress Validity Findings Targeting the Content-Focused Vocabulary Matching. *Exceptional Children*, *80* (1), 85–100.

Morse, D. T. Review of the Stanford Achievement Test, 10th ed. *Mental Measurements Yearbook*. Prieiga internete: http://buros.org/mental-measurements-yearbook

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for English Language Arts and Literacy in History/Social Studies, Science, and Technical Subjects*. Washington, DC: Authors.

*Pearson Education*. *Stanford Achievement Test Series, online abbreviated form* (10th ed.). Prieiga internete: http://www.pearsonassessments.com/learningassessments/products/100000563/stanford-achievement-test-series-tenth-edition-abbreviated-battery.html

Royer, J. M. (2004). Uses for the Sentence Verification Technique for Measuring Language Comprehension. *Progress in Education*. Prieiga internete: http://www.readingsuccesslab.com/publications/Svt%20Review%20PDF%20version.pdf

Royer, J. M., Hastings, C. N., & Hook, C. (1979). A Sentence Verification Technique for Measuring Reading Comprehension. *Journal of Reading Behavior*, *11*, 355–363.

Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using Curriculum-Based Measurement to Improve Student Achievement: Review of Research. *Psychology in the Schools*, *42*, 795–819. doi: 10.1002/pits.20113.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using Multivariate Statistics* (6th ed.). Boston, MA: Pearson Education, Inc.

Tolar, T. D., Barth, A. E., Fletcher, J. M., Francis, D. J., & Vaughn, S. (2014). Predicting Reading Outcomes with Progress Monitoring Slopes among Middle Gra-

de Students. *Learning and Individual Differences*, *30*, 46–57. Doi:10.1016/j.lindif.2013.11.001.

Vannest, K. J., Parker, R., & Dyer, N. (2011). Progress Monitoring in Grade 5 Science for Low Achievers. *The Journal of Special Education*, *44*, 221–233.

Wallace, T., Espin, C. A., McMaster, K., Deno, S. L., & Foegen, A. (2007). CBM Progress Monitoring within a Standards-Based System: Introduction to the Series. *The Journal of Special Education*, *41*, 66–67.

Warne, R. T. (2011). Beyond Multiple Regression Using Commonality Analysis to Better Understand R 2 Results. *Gifted Child Quarterly*, *55* (4), 313–318.

Zientek, L. R., & Thompson, B. (2006). Commonality Analysis: Partitioning Variance to Facilitate Better Understanding of Data. *Journal of Early Intervention*, *28* (4), 299–307.

## BENDRŲJŲ GAMTOS PAŽINIMO IR SOCIALINIŲ MOKSLŲ MOKYMOSI REZULTATŲ VERTINIMO VALIDUMAS

Paul Mooney, Renée E. Lastrapes, Amanda M. Marcotte,
Amy Matthews, B. S.

### Santrauka

Mokinių skaitymo ir matematikos gebėjimų formuojamasis vertinimas ilgą laiką buvo specialiojo ugdymo sistemos dalis. Tyrimų, skirtų struktūruoto formuojamojo vertinimo technologijoms analizuoti, siekiant užfiksuoti mokinio pasiekimus ir augimą mokantis socialinių mokslų ir gamtos pažinimo, vis dar nėra daug. Pirmiausia buvo vertinami mokinių skaitymo įgūdžiai, siekiant nustatyti jų koreliaciją su dalyko turinio suvokimu. Pastaruosius du dešimtmečius literatūros šia tema padaugėjo, ji apėmė į turinį orientuotus tyrimo instrumentus.

Tyrimas analizavo ir kitus tris į turinį orientuotus instrumentus, kurie yra pažangesni, nes administruojami ir įvertinami internetu: dalyko turinio suvokimą, teisingų ar klaidingų teiginių žymėjimą ir rašytinį atpasakojimą. Šie instrumentai buvo vertinami už jų techninį adekvatumą ir logistinį įgyvendinamumą. Straipsnyje analizuojami ir du tyrimo klausimai: (a) koreliacija su nacionaliniais norminiais standartizuotais pasiekimų bei valstijos atsiskaitomaisiais

testais gamtos pažinimo ir socialinių mokslų turinio mokymosi sričių turinio suvokimo, teisingų ar klaidingų teiginių žymėjimo ir rašytinio atpasakojimo atvejais ir (b) validumo augimas, papildant teksto supratimo vertinimą teisingų ar klaidingų teiginių žymėjimo metodu ir (ar) rašytiniu atpasakojimu turiniu grįstuose pasiekimų modeliuose, kuriais naudojantis buvo tiriamas dalyko turinio suvokimas.

Tyrimo dalyviai buvo penktos klasės mokiniai (N = 51), lankantys valstybinę pradinę mokyklą pietrytinėje JAV dalyje. Šie dalyviai buvo 11,1 metų amžiaus (standartinis nuokrypis SD = ,5) testavimo metu, 68,6 proc. mergaičių (n = 35), 66,7 proc. baltaodžių (n = 34) ir 68,6 proc. gaunančiųjų nemokamus pietus (n = 35). Trys numatyti kintamieji (t. y. turinio suvokimas, teiginių žymėjimo metodas ir rašytinis atpasakojimas) koreliavo su turinio testo rezultato balais, gautais iš nacionalinio reprezentuojamojo standartizuoto pasiekimų testo (t. y. Stanfordo pasiekimų testo dešimtosios laidos sutrumpinta internetinė forma) ir valstijos atsiskaitomojo testo. Pirsono (Pearson) koreliacija tarp turinio suvokimo ir Stanfordo gamtos pažinimo (r = ,55) ir socialinių mokslų (r = ,63) sričių testų buvo vidutiniškai stipri ir dydžiu panaši į kitas koreliacijas akademinės kalbos tyrimo atvejais. Turinio suvokimo koreliacija buvo didesnė nei nustatytosios tarp standartizuotų testų ir teisingų ar klaidingų teiginių žymėjimo metodo bei rašytinio atpasakojimo. Panašumų analizė rodo, kad tiek turinio suvokimas, tiek teisingų ar klaidingų teiginių žymėjimo metodas aiškinamuosius modelius papildė unikaliais skirtumais.

Rezultatų aptarimas atskleidė, kad, pirma, akademinės kalbos tyrimas yra perspektyvus dalykas, turint omenyje struktūruoto formuojamojo vertinimo instrumentų, pavyzdžiui, žodyno atitikties ir turinio suvokimo, taikymą tolesniuose tyrimuose; antra, edukacinės rekomendacijos yra grindžiamos daugeliu šaltinių; šio tyrimo rezultatai teigia, kad daugelis instrumentų prisideda prie aiškinamųjų modelių unikalios įvairovės atsiradimo ir kad pasiteisina būtent daugialypio struktūruoto formuojamojo vertinimo matavimų taikymas, plėtojant turinio suvokimo vertinimo schemas.

Tyrimo vertę šiek tiek mažina instrumentų pristatymo eiliškumas, mažas imties dydis ir jos sudarymas bei koncentravimasis į vienos klasės lygį.

doi: 10.21277/se.v1i34.253

# VALIDITY OF TWO GENERAL OUTCOME MEASURES OF SCIENCE AND SOCIAL STUDIES ACHIEVEMENT

Paul Mooney
Special Education Programs
Louisiana State University
213 Peabody Hall, USA

Renée E. Lastrapes
University of Houston
UH-Clear Lake Box 55
2700 Bay Area Blvd. Houston, Texas 77058

Amanda M. Marcotte
University of Massachusetts Amherst
S154 Furcolo
Amherst, MA 01003

Amy Matthews, B. S.
Louisiana State University
13046 LA-73, Geismar, LA 70734

**Abstract**

The present research expanded validity findings for a structured formative assessment measure of content learning that was administered online and known as critical content monitoring. The study also evaluated the potential for additional measures, including sentence verification technique and written retell, to explain variance in student achievement in science and social studies classrooms. Participants were fifth-grade students (N=51) enrolled in a public primary school in the southeastern U.S. Three predictor variables (i.e. critical content monitoring, sentence verification technique and written retell) were correlated with content test scores from the nationally representative standardized achievement test (i.e. Stanford Achievement Test-Tenth Edition abbreviated online form) and a statewide accountability test. Pearson correlations for critical content monitoring and the Stanford tests across science (r=.55) and social studies (r=.63) were moderately strong and similar in magnitude with other reported correlations for academic language measures in the literature. Correlations for critical content monitoring were descriptively larger than those between the standardized tests and sentence verification technique and written retell. Commonality

analyses indicated that both critical content monitoring and sentence verification technique added unique variance to explanatory models. Limitations and implications were discussed.

**Keywords:** *structured formative assessment, general outcome measurement, content courses.*

Structured formative assessment has long been part of the fabric of special education, both in the profession's practice and promise (Jenkins, & Fuchs, 2012). Since Dr. Stanley Deno's development and introduction of curriculum-based measurement (Deno, 1985), special education teachers have had the capacity to monitor their students' progress toward individualized education programme goals. Curriculum-based measurement has been implemented and evaluated primarily in the elementary grades and in the area of reading skill development. However, from its inception, assessment procedures have also been available in the areas of beginning math, spelling, and writing skills. What did not exist initially were procedures to evaluate student performance and progress in content areas such as social studies and science. It is the content areas to which this research applies.

Inquiry addressing the efficacy of structured formative assessment techniques to document student achievement and growth in social studies and science content is still in its infancy. Early on, researchers such as Espin and Foegen (1996) wondered whether available reading measures such as oral reading fluency and maze could also serve to document performance in content areas. They found that a measure of content vocabulary was a descriptively stronger correlate than the reading measures. From that beginning has evolved a more focused evaluation of academic content-driven measures, including tests of the validity of critical content monitoring (Mooney, McCarter, Russo, & Blackwood, 2013), an online-administered and scored structured formative assessment tool, that is the focus of the present study. What follows is a description of critical content monitoring and a rationale for the following two research questions:

1. What were the correlations with nationally-normed standardized achievement and statewide accountability tests in science and social studies content for critical content monitoring, sentence verification technique and written retell?

2. What was the incremental validity of adding measures of reading comprehension, using sentence verification technique and/or written retell, to content-focused achievement models that included critical content monitoring?

## Critical content monitoring

Critical content monitoring was originally developed using a curriculum sampling approach (Fuchs, 2004). Probes were created by sampling the corpus of content vocabulary across a grade-level curriculum in order to ensure that all important content was included across all forms. As a general outcome measure, its aim has been to serve as an index of content learning at both a point in time and over time in order to operate formatively and improve instructional decision-making. Administration of critical content monitoring probes involves students reading definitions of key grade-level science or social studies vocabulary at a secure learning management system link and choosing the correct answer from a list of terms. Students generally have up to 5 minutes to answer 20 questions.

Academic vocabulary was utilized as an indicator of performance because it serves as communicative currency (Alexander, n.d.) in content courses. That is, the words, phrases, and concepts of the subject matter form the content of activities in the classroom. Success in the content classroom occurs through relevant and meaningful employment of academic vocabulary. Demonstration of academic language's robustness is evident in the stronger correlations with a relevant criterion for vocabulary matching over competing measures including oral reading fluency, maze and a writing measure (Espin, & Foegen, 1996; Mooney, McCarter, Schraven, & Callicoatte, 2013).

Critical content monitoring is an online adaptation of vocabulary matching (Espin & Deno, 1994-1995). Its research has addressed technical concerns of a single probe score. Mooney, McCarter, Russo et al. (2013) assessed the criterion validity for a collection of 20 science probes in relation to a statewide accountability test for a sample of generally high-performing fifth-grade students. The results indicated moderate correlations for the 20 probes (r=.36 - .55). Mooney, McCarter, Russo, Blackwood (2014) extended the criterion validity findings, demonstrating a moderately strong correlation (.67) between a critical content monitoring social studies probe and the statewide content test. The social studies correlation was descriptively larger than the science correlation and comparable in magnitude to previous vocabulary matching findings (rs .64 to .70; Espin, Shin, & Busch, 2005; Mooney, McCarter, & Schraven et al., 2013).

## Rationale for study

The predictor measures in the present study were either originally designed as general outcome measures of learning or more recently adapted for that purpose. General outcome measurement (GOM) is one of two instructionally relevant measurement models (Fuchs & Deno, 1991). An alternative to subskill mastery measurement, its original goal was to establish tests that: (a) evidenced reliability and validity, and (b) assisted teachers in planning better instructional programmes and evaluating instructional programme success. These findings enhance the larger literature in content-oriented structured formative assessment, including studies that have addressed the performance and progress tenability of vocabulary matching (e.g., Espin, Lembke, Hampton, Seo, & Zukowski, 2013; Mooney, McCarter, & Schraven et al., 2013).

General outcome measures of academic language, including critical content monitoring, vocabulary matching and key vocabulary (Vannest, Parker, & Dyer, 2011) have the potential to predict achievement and inform instructional decision-making. Critical content monitoring, for example, was intended as a measure of science or social studies course learning by content teachers in the upper elementary and secondary school grades. For teachers, academic language is an alterable variable (Bloom, 1980) that is particularly pertinent to their interaction with struggling learners. Still, more research is warranted as GOM-type assessment tools are created to reflect broader academic domains. Sentence verification technique and written retell were evaluated as potential general outcome measures of reading comprehension by Marcotte and Hintze (2009). General outcome measures, particularly in reading, have been demonstrated to be effective tools to predict student achievement and inform teacher decision-making (Stecker, Fuchs, & Fuchs, 2005; Wallace, Espin, McMaster, Deno, & Foegen, 2007). Moreover, they have been instrumental in the implementation of responsiveness-to-intervention (RTI) frameworks.

The aforementioned research questions relate to what Fuchs (2004) termed Stage 1 viability of the static score, and Deno and Fuchs (1987) categorized as technical adequacy questions in their instrument development matrix. The first question addressed the need to extend validity research for critical content monitoring. The measure has validity findings comparing scores with a statewide accountability test across both science and social studies content (Mooney, McCarter, & Russo et al., 2013, 2014). The first research question extended criterion validity evidence by comparing critical content monitoring scores with those of a standardized measure of content knowledge. Generalization concerns exist because critical monitoring comparisons to date have been made with state-specific accountability tests. The present inquiry was the first

comparison between scores of critical content monitoring and a standardized, nationally-recognized test of science and social studies achievement. It was hypothesized that correlations with a nationally-representative test would be comparable to those with statewide accountability tests.

The second research question addressed the complexity of knowledge and skills that may be relevant for measurement in a formative assessment model. Indicators of academic language, such as vocabulary matching, critical content monitoring and key vocabulary, are not the only variables that predict criterion achievement so other indicators of success in content areas should be evaluated. That is, while students are expected to master academic language, they are also expected to read and comprehend instructional texts and relevant materials presented in class (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010) as well as summarize and apply what they learn and read in written form.

Marcotte and Hintze (2009) evaluated a series of reading comprehension explanatory models that paired oral reading fluency with a number of potential GOM comprehension tools, including maze, retell fluency, sentence verification technique and written retell. Correlations for all of the GOM predictors with a criterion were moderately strong, ranging from .46 to .67. Multiple-regression analyses indicated that a collection of four of the measures (i.e. oral reading fluency, maze, sentence verification technique and written retell) accounted for about 57% of the variability in overall reading ability. Moreover, the maze, sentence verification technique and written retell measures were all found to contribute significantly to the overall explanatory model after controlling for oral reading fluency. One takeaway from the study was that four measures together proved to be a better predictive mechanism than one measure alone.

In the present study three measures targeting academic language and comprehension were evaluated to determine whether there would be a greater share of the variability in content achievement for multiple measures of content achievement beyond that provided for by critical content monitoring. Potential GOM tools, sentence verification technique and written retell, were chosen over oral reading fluency and maze because of their potential value in describing variance in content achievement. Moreover, previous research demonstrated that a measure of academic language was the more strongly correlated instrument with a criterion when comparisons were made either descriptively or directly (Espin, Foegen, 1996; Mooney, McCarter, & Schraven et al., 2013), making it a preferable capstone task. It was hypothesized that a multi-element assessment system would be a better predictor than a single-measure system.

## Method

### Participants and setting

Participants in the institutional review board-approved study were fifth-graders (N = 51) in a single public prekindergarten to fifth grade school in south Louisiana whose parents consented and who they assented to be involved in the study. As a whole, participants were 11.1 years old (SD = .5) at the time of testing, 68.6% female (n = 35), 66.7% Caucasian (n = 34), and 68.6% full-pay lunch status (n = 35). 96% of the participants (n = 49 in both cases) scored at the 'basic' level of proficiency on the statewide accountability content tests (which would be considered passing if science and social studies test scores were categorized in the same manner as the English language arts and math tests were in Louisiana). Statewide, 43% and 47% of fifth-graders scored at 'basic' in science and social studies, respectively, on the state test. Assenting participants were 40.8% of an entire fifth-grade class.

### Measures

Five measures were compared in the present study. The criterion measures were the science and social studies content tests of the online abbreviated Stanford Achievement Tests-Tenth Edition (SAT-10; Pearson Education, n.d.) and the *integrated* Louisiana Educational Assessment Programme (*i*LEAP; LDE, n.d., a). The predictor variables were critical content monitoring, sentence verification technique and written retell.

**SAT-10.** The abbreviated form of the online SAT-10 is a standardized, norm-referenced achievement test battery that measures reading, mathematics, spelling, language, listening, science and social studies performance for students in kindergarten through the 12$^{th}$ grade. For this study, only grade-level science and social studies tests were administered. Publishers described the content tests as reflecting current practice and research and aligned with national and state content standards. The science test assessed knowledge of life, physical and earth sciences as well as science as inquiry. The social studies test assessed knowledge of history, geography, political science and economics. Each abbreviated battery content test consisted of 30 multiple-choice questions and was untimed. The test-derived scaled score was used in the present study. The scaled score is vertically equated across each subject test, reportedly allowing for the tracking of performance across grades (Pearson Education, n.d.). For the Spring 2013 testing period, the sample's average science score was 658 (range 602-726), placing the collective group at the 63$^{rd}$ percentile nationally. The

average social studies score of 655 (range 597-713) was reported at the 59th percentile. Two online Buros Institute *Mental Measurements Yearbook* reviewers (Carney, n.d.; Morse, n.d.) provided support for the use of SAT-10 as a whole in measuring achievement in public schools. Reviewers described alternate-form reliability and content validity evidence for SAT-10. In the present study, criterion validity was evidenced by moderately strong correlations between the SAT-10 and *i*LEAP content tests, with .64 and .69 linear relations, both $p < .01$, for science and social studies, respectively.

***i*LEAP grade 5 criterion-referenced test.** The stated purpose of *i*LEAP is measurement toward Louisiana's academic standards in English language arts, math, science and social studies (LDE, n.d., a) for all students in grades 3, 5, 6, 7 and 9. The science and social studies tests included multiple-choice questions, were untimed and administered on different days. Fifth grade science content strands included science as inquiry, physical, life, earth, space and environmental science, with test questions addressing all strands. Social studies content strands included geography, civics, economics and history, with test questions addressing only the geography and history strands (LDE, n.d., a). Achievement level descriptors were unsatisfactory, approaching basic, basic, mastery and advanced. Technical adequacy data for the *i*LEAP fifth grade tests were accessed from the LDE website. Cronbach's alpha levels of .85 for science and .82 for social studies were reported as reliability evidence of the 2010 test's internal consistency (LDE, n.d., b). State-provided validity data were described in terms of a content validity process that was not delineated (LDE, n.d., b).

**Critical content monitoring**. The content-focused general outcome measure described earlier evolved from procedures previously outlined in Espin, Busch, Shin, Kruschwitz (2001). Terms in each researcher-created probe were randomly selected from the full body of content terms collected from textbook glossary sections and reviewed for legitimacy by a small group of practicing teachers recommended by the first author as both content knowledgeable and effective teachers. The list of terms was organized by curricular unit. Each probe included terms from each unit. To generate each probe, the number of terms per unit was determined by calculating the proportion of the year's curriculum that was devoted to each unit in the state pacing guide and then multiplying that proportion by the number of questions in each probe. Twenty terms and accompanying definitions were entered into an online learning management system (Moodle, n.d.) in a multiple-choice format. Alternate form reliability correlations for 20 parallel probes had a mean correlation of .55 (range .21 to .73; SD = .09) (Mooney, McCarter, & Russo et al., 2013). Criterion validity findings were previously reported.

**Sentence verification technique**. Sentence verification technique was reportedly created as a reading comprehension assessment method (Royer et al., 1979). The measure consisted of reading passages that were followed by sentences that test takers read and responded to after reading the passage and without access to the passage content. Sentences were developed from the passages that test takers read and were one of the following type: (a) originals, or exact copies of passage sentences; (b) paraphrases, or similar-meaning sentences with built-in word changes; (c) meaning changes, or similar sentences with slight changes in words to alter meaning; and (d) distractors, or similar topic sentences that differ in both meaning and wording from the passage. For the present study, passage content was adapted from approved grade-level science and social studies texts. Passages were split between science and social studies content and introduced alternately, with 16 sentences following each passage. An examinee read each passage and then responded to the sentences with a yes (i.e. the meaning of the sentence is the same as that of the passage) or no (i.e. the meaning is different) response. Criterion validity correlations with standardized test measures, including the SAT, have ranged from .50 to .73; reliability for four-passage probes have ranged from .70 to .80 (Royer, 2004).

**Written retell.** The format for written retell was identical to that utilized in Marcotte and Hintze (2009). That is, students were asked to read a 750-word passage silently for 5 minutes. Then, the passage was removed and students were asked to write down all that they could remember about the passage. Students had 5 minutes to respond to the initial prompt, which included reminders of the task requirement offered periodically during that time span. Scores for written retell consisted of the number of unique content words written by each student. As in Marcotte and Hintze, content words were defined as "distinct proper and common nouns, verbs, adjectives, and adverbs contained in the passage or synonymous with those in the passage" (p. 322). The list of content words was developed by the first author based upon a reading of the passage. A criterion validity correlation for written retell with a standardized achievement test was reported at .57 (Marcotte & Hintze). The written retell probe also demonstrated .56 and .59 correlations with oral reading fluency and maze, respectively.

## Procedures

Testing took place during science class in mid-May 2013, near the end of the school year and about six weeks after statewide accountability testing. Originally, test administration was designed in a counter-balanced arrangement by class section in order to address order effects. However, technical difficulties

encountered on the first occasion of SAT-10 online testing resulted in an alteration of the original schedule, with SAT-10 testing taking place last for all participants. For SAT-10 and critical content monitoring, students logged on to secure sites under the supervision of the first author. For sentence verification technique and written retell, testing was directed by the first and fourth authors, with the latter the classroom science teacher. For the statewide accountability tests, the classroom teacher oversaw administration.

## Interscorer agreement

Sentence verification technique and written retell were independently scored by the first and second authors. The two scores were entered into a database and checked to ensure accurate data entry. Initial data analyses included the calculation of agreement proportions for all participant total scores. An agreement occurred when both scorers reported the same total score per individual probe. Agreement proportions were 100% for sentence verification technique and 85.7% for written retell. The first scores from the first author were used for the analyses. No interscorer reliability actions were taken for critical content monitoring beyond checking the online system to ensure that the right answer choice accompanied each stem. SAT-10 scores were provided by the publishing company whereas *i*LEAP scores were provided by the teacher.

## Analysis

Analyses addressed criterion and incremental validity questions. Criterion validity for each of the three predictor variables was assessed using the correlation and the 95% confidence interval (CI) between each probe and a relevant criterion measure. All variables were assessed and deemed normally distributed using the Shapiro-Wilk statistic. Incremental validity was assessed using sequential multiple linear regression and commonality analysis. Separate sequential or hierarchical multiple regression analyses were performed for science and social studies content areas. Sequential regression was used to determine if information from additional predictor variables improved prediction of the criterion measure after the effects of previous variables entered had been statistically eliminated (Tabachnick & Fidell, 2013). In order to determine how much more variance was accounted for by each of the comprehension measures critical content monitoring was entered first, followed by sentence verification technique and then written retell. Finally, commonality analyses were conducted to determine the amount of variation in the criterion

variables (i.e. SAT-10 science and social studies) accounted for by each of the predictor variables. Commonality analysis is designed to quantify the unique explanatory power of each predictor as well as the explanatory power that is common to all possible combinations of the predictors (Zientek & Thompson, 2006). The commonality formulas for Excel were obtained from Warne (2011).

## Results

### Criterion validity

Table 1 displays sample means, distributions and 95% CIs of state and national content tests and each of the three predictors. While skewness and kurtosis were evident for the predictors, normality checks showed that all but the critical content monitoring science scores were normally distributed. Table 2 displays correlations of all assessments with 95% CIs. Correlations and 95% CIs between the predictor and criterion variables were in the low (i.e. <.3) to moderate (i.e. .3 to .7) range. Critical content monitoring was most highly correlated with both content tests, followed next by sentence verification technique and then written retell. Social studies correlations were descriptively greater in magnitude than those in science. Each of the correlations for critical content monitoring and sentence verification technique and the criterion measures were significantly different from zero ($p \leq .01$). Sentence verification technique showed stronger correlations with the social studies tests than science, and WRT was only significantly correlated with the state social studies test.

Table 1

**Means of State Accountability Tests, Standardized Tests, and Predictor Tests in Fifth Grade**

| | N | Range | Mean | SD | 95% CI | Skew-ness | Kurto-sis |
|---|---|---|---|---|---|---|---|
| *i*LEAP Science | 51 | 276-454 | 360.6 | 36.8 | 350, 371 | .33 | .62 |
| *i*LEAP Social Studies | 51 | 281-424 | 347.3 | 31.7 | 338, 356 | .29 | -.01 |
| SAT-10 Science | 51 | 602-726 | 658.5 | 25.6 | 651, 666 | .15 | .29 |
| SAT-10 Social Studies | 46 | 586-713 | 654.0 | 27.7 | 646, 662 | .06 | .17 |
| CCM Science | 49 | 8-20 | 15.73 | 3.05 | 14.86, 16.61 | -0.80 | 0.02 |
| CCM Social Studies | 50 | 3-18 | 11.34 | 4.34 | 10.08, 12.60 | -0.08 | -1.08 |
| SVT | 50 | 29-59 | 47.36 | 6.54 | 45.50, 49.22 | -0.32 | 0.34 |
| WRT | 49 | 11-42 | 26.92 | 7.86 | 24.66, 29.18 | 0.39 | -0.78 |

*Note.* CCM = critical content monitoring; SVT = sentence verification technique; WRT = written retell; *i*LEAP = *integrated* Louisiana Educational Assessment Program; SAT-10 = Stanford Achievement Test-Tenth Edition, abbreviated form.

Table 2

**Correlations and 95% Confidence Intervals Among all Predictor and Criterion Measures**

| Test | *i*LEAP Social Studies | SAT Science | SAT Social Studies | CCM Science | CCM Social Studies | SVT | WRT |
|---|---|---|---|---|---|---|---|
| *i*LEAP Science | .66** [.47, .79] | .64** [.44, .78] | .57** [.35, .73] | .51** [.27, .69] | .47** [.22, .66] | .46** [.21, .65] | .24 [-.03, .48] |
| *i*LEAP Social Studies | | .49** [.25, .68] | .69** [.54, .81] | .61** [.40, .76] | .66** [.47, .79] | .49** [.25, .68] | .29* [.02, .52] |
| SAT Science | | | .51** [.27, .69] | .55** [.32, .72] | .43** [.18, .63] | .49** [.25, .68] | .16 [-.12, .42] |
| SAT Social Studies | | | | .65** [.26, .70] | .63** [.42, .77] | .59** [.36, .75] | .28 [-.01, .53] |
| CCM Science | | | | | .60** [.38, .75] | .40** [.13, .61] | .26 [-.02, .50] |
| CCM Social Studies | | | | | | .41** [.15, .62] | .25 [-.03, .49] |
| SVT | | | | | | | .22 [-.06, .54] |

*Note.* ** Correlation is significant at the .01; * correlation is significant at .05 level. CCM = critical content monitoring; SVT = sentence verification technique; WRT = written retell; *i*LEAP = *integrated* Louisiana Educational Assessment Program; SAT = Stanford Achievement Test-Tenth Edition, abbreviated form.

## Incremental validity

Based on the results of the correlational analysis, the predictors were entered into the sequential regression model with critical content monitoring first, sentence verification technique second and written retell third. Data were analyzed for influential points and multicollinearity. Because two observations showed values of Cook's D that were greater than 1, regression analyses were conducted with and without them and results did not change appreciably; therefore, the observations were retained in the data set. Formal tests for multicollinearity revealed that none was detected. Residual plots indicated homoscedasticity as well as linear relationships among all predictor and criterion variables.

Tables 3 and 4 display results of the regression analyses for the three predictor variables and the criterion SAT-10. Findings indicated that the greatest amount of variation was associated with critical content monitoring and sentence verification technique. Written retell was not a significant predictor either science or social studies, with adjusted $R^2$ decreasing for both SAT-10 test scores with its inclusion.

Table 3

**Sequential Regression for SAT-10 Science with Predictors Entered in the Order Presented**

| | | | | | Model 1 | | Model 2 | | Model 3 | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Predictors** | **R** | **$R^2$** | **Adj. $R^2$** | **$R^2$ change** | **β** | **p value** | **β** | **p value** | **β** | **p value** |
| CCM | 0.54 | 0.30 | 0.28 | 0.30 | 4.83 | .000 | 3.75 | .002 | 3.79 | .002 |
| CCM, SVT | 0.61 | 0.37 | 0.35 | 0.08 | | | 1.21 | .023 | 1.22 | .024 |
| CCM, SVT, WRT | 0.61 | 0.37 | 0.33 | 0.00 | | | | | -0.08 | .855 |

*Note.* SAT-10 = Stanford Achievement Test-Tenth Edition, abbreviated form; CCM = critical content monitoring; SVT = sentence verification technique; WRT = written retell.

Table 4

**Sequential Regression for SAT-10 Social Studies with Predictors Entered in the Order Presented**

| | | | | | Model 1 | | Model 2 | | Model 3 | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Predictors** | **R** | **$R^2$** | **Adj. $R^2$** | **$R^2$ change** | **β** | **p value** | **β** | **p value** | **β** | **p value** |
| CCM | 0.62 | 0.38 | 0.36 | 0.38 | 3.64 | .000 | 2.73 | .000 | 2.65 | .000 |
| CCM, SVT | 0.72 | 0.52 | 0.50 | 0.14 | | | 1.76 | .001 | 1.72 | .001 |
| CCM, SVT, WRT | 0.73 | 0.53 | 0.49 | 0.01 | | | | | 0.23 | .452 |

*Note.* SAT-10 = Stanford Achievement Test-Tenth Edition, abbreviated form; CCM = critical content monitoring; SVT = sentence verification technique; WRT = written retell.

Table 5 displays results of the commonality analyses. For science the regression effect was most influenced by critical content monitoring (14.8%) and sentence verification technique (7.3%). There was no measurable variance that was uniquely attributable to written retell. When examined in combination, critical content monitoring and sentence verification technique accounted for a third of the variation in the criterion variable (14.8 + 7.3 + 12.8 = 34.9%).

When examined in combination with critical content monitoring, written retell appeared to have a suppressor effect on critical content monitoring, as evidenced by the negative coefficient. Table 5 also indicates that for SAT-10 social studies scores, the regression effect was most influenced uniquely by critical content monitoring (17.3%) and sentence verification technique (11.3%). The variance that was unique to written retell again appeared to act as a suppressor variable (-1.6% of the total). When examined in combination, critical content monitoring and sentence verification technique accounted for 44.9% of the variation in SAT-10 social studies (17.3 + 11.3 + 16.3).

Table 5

**Commonality Analyses for SAT-10 Science and Social Studies**

| SAT Science | CCM | SVT | WRT | $R^2$ Partition | SAT Social Studies | CCM | SVT | WRT | $R^2$ Partition |
|---|---|---|---|---|---|---|---|---|---|
| $U_1$ | 14.8% | -- | -- | 14.8% | $U_1$ | 17.3% | -- | -- | 17.3% |
| $U_2$ | -- | 7.3% | -- | 7.3% | $U_2$ | -- | 11.3% | -- | 11.3% |
| $U_3$ | -- | -- | 0.0% | 0.0% | $U_3$ | -- | -- | -1.6% | -1.6% |
| $C_{12}$ | 12.8% | 12.8% | -- | 12.8% | $C_{12}$ | 16.3% | 16.3% | -- | 16.3% |
| $C_{13}$ | -1.0% | -- | -1.0% | -1.0% | $C_{13}$ | 1.7% | -- | 1.7% | 1.7% |
| $C_{23}$ | -- | 0.1% | 0.1% | 0.1% | $C_{23}$ | -- | 3.1% | 3.1% | 3.1% |
| $C_{123}$ | 3.4% | 3.4% | 3.4% | 3.4% | $C_{123}$ | 4.4% | 4.4% | 4.4% | 4.4% |
| Sum = $r^2$ | 30.0% | 23.6% | 2.5% | -- | | 39.7% | 35.1% | 7.6% | -- |
| Sum = $R^2$ | | | | 37.4% | | | | | 52.5% |
| Unique | 14.8% | 7.3% | 0.0% | | Unique | 17.3% | 11.3% | -1.6% | |
| Common | 15.2% | 16.3% | 2.5% | | Common | 22.4% | 23.8% | 9.2% | |
| Total | 30.0% | 23.6% | 2.5% | | Total | 39.7% | 35.1% | 7.6% | |

*Note.* SAT-10 = Stanford Achievement Test-Tenth Edition, abbreviated form; CCM = critical content monitoring; SVT = sentence verification technique; WRT = written retell. U = Unique; C = Common; 1 = CCM; 2 = SVT; 3 = WRT.

## Discussion

Providing evidence of test score validity is a vital consideration in test development (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999). It is also an important notion for special educators who are charged with individualizing educational programming and ensuring that those unique practices are based on research to the extent practicable. The present study

offered further evidence of criterion validity for critical content monitoring as well as initial evidence of validity for a package of potential GOM instruments for formative use in public school science and social studies classrooms.

Critical content monitoring scores correlated moderately in magnitude (i.e. .3 to .7) in science and social studies to scores from the nationally-recognized SAT-10, which was expected and necessary for consideration of the generalizability of these findings. The science correlation of .55 (95% CI; .32, .72) was within the .36 to .55 range for linear relations with a statewide test previously reported in Mooney, McCarter, Russo et al. (2013). The moderate magnitude of correlation between critical content monitoring and state level accountability test (i.e. .51) was replicated as well (see Table 2). The social studies correlation with SAT-10 of .63 (95% CI; .42, .77) was descriptively higher than the science correlation and comparable to the .68 linear relation with a state accountability test reported in Mooney et al. (2014) and .66 in this study (see Table 2). A pattern of critical content monitoring relations that were descriptively higher in social studies than in science was maintained across national and state assessments. Further evidence of the validity of the critical content monitoring score was indicated by its explanatory power. In regression models identifying the amount of variance accounted for by predictor variables, critical content monitoring was the strongest predictor for both science and social studies tests of the SAT-10. Moreover, commonality analyses indicated that 15% to 17% of the variance in models was uniquely explained by critical content monitoring.

Findings from the present study also identified sentence verification technique as a potential indicator of student performance in the content area courses, with 7% to 11% of the variance explained by sentence verification technique alone and roughly 35% to 45% explained by critical content monitoring and sentence verification technique combined. Written retell was not a significant contributor to the regression effect and showed little common variance with the criterion variable. Both previously had been evaluated as potential measures of reading comprehension by Marcotte and Hintze (2009). For sentence verification technique, which was previously found to have moderate correlations with criteria and explanatory power in models of reading comprehension (Marcotte, Hintze), scores were also moderately correlated with science and social studies content test scores of the SAT-10, with linear relations of .49 (95% CI; .25 .67) in science and .59 (95% CI; .37, .75) in social studies. These results provide evidence that critical content vocabulary and reading comprehension are meaningful predictors of achievement in science and social studies as measured by both state and nationally represented achievement tests

with sentence verification technique a strong potential candidate as an effective general outcome measure of content knowledge.

For written retell, neither of its correlations with the SAT-10 was moderate in magnitude or statistically significant in this study. Findings contrast those of Marcotte and Hintze (2009), who found that written retell was a significant predictor of models of reading comprehension. The disparate findings for written retell may have resulted from the fact that written retell is scored by the number of meaningful words recalled and written from the passage, which in the context of science, students may have less familiarity with the vocabulary and thus have been less likely to recall them and write them down.

## Limitations

Three primary limitations were noted in the present study. First, due to technical difficulties during the original implementation of the research plan, the order of testing was altered and SAT-10 testing took place last for all students, potentially opening results up to the influence of order effects. Second, the participant pool was a convenient sample of assenting, high-performing individuals and may not be representative of the larger public school population. When compared to national samples, the group average for participants in this sample was at the 63$^{rd}$ and 59$^{th}$ percentiles in SAT-10 science and social studies, respectively. Third, with the sample consisting of students from a single grade, caution is warranted when generalizing results to other grade levels.

## Implications

With the present limitations and findings described, it is noteworthy that evidence continues to support consideration of indicators of academic language as viable GOM tools in the documentation of performance and progress in science and social studies content. While still focused on Stage 1 (of 3) evaluation of the static score (Fuchs, 2004), the present findings lay a foundation for a technically adequate, instructionally effective and logistically feasible assessment framework in the content areas that include general outcome measures of academic language and comprehension.

A first implication relates to the utility of academic vocabulary as a measurement index in the content areas. A growing body of comparison research in the GOM of content areas favor instruments like critical content monitoring, vocabulary matching, and key vocabulary. Espin and Foegen (1996)

made descriptive comparisons of three predictors (i.e. vocabulary matching, maze, and oral reading fluency) and reported that the largest correlations with the criterion were primarily those with vocabulary matching. Mooney, McCarter, Schraven et al. (2013) made direct comparisons between the correlations of vocabulary matching, maze, and a writing GOM and found statistically significant differences in correlations with a statewide accountability test that favored vocabulary matching. Multiple regression analyses in these studies provided evidence that the measure of academic language provided the strongest explanatory power in the series of models analyzed. That pattern was repeated with critical content monitoring in the present study.

Furthermore, the utility of GOM with academic language as the index appears to extend to both science and social studies content areas. Initial research in this area focused on social studies learning, with linear relations with criterion measures across the upper elementary and middle school grades generally reported to be in the moderate to strong range (e.g., Espin et al., 2001). However, recent studies, including this one, have extended the evaluation of linear relationship to science across the elementary and secondary grades (e.g., Espin et al., 2013), with findings that have generally been moderate in magnitude and statistically significant, ranging from .45 to .66. The fact that a GOM index of content learning can be applied across subject areas bolsters the potential of the tool to be instructionally effective, something critical to RTI assessment frameworks. Moreover, logistical feasibility is apparent in the fact that the test can be administered using online technology. Such findings provide special educators with a research-informed choice when it comes to measuring academic performance in science and social studies classrooms.

A second implication relates to continued discovery of appropriate measurement technologies for use in upper elementary and secondary settings. The uses of GOM continue to grow across the K-12 spectrum. General outcome tools continue to be developed and evaluated as do assessment systems. Recently, Deno and colleagues (2009) described the implementation of a schoolwide elementary grades reading assessment program that incorporated universal screening and progress monitoring using maze and oral reading fluency measures in a manner that attempted to document progress across grades and over time. Similar efforts to develop long-term and sensitive assessment technologies in the content areas appear warranted.

Multiple measures appear worthy of continued exploration. Across science and social studies content, critical content monitoring demonstrated

the greatest explanatory power in regression models. Sentence verification technique, a measure of reading and possibly content comprehension, also was a significant predictor in both content areas. Commonality analyses (see Table 5) indicated that both measures had their own and shared substantial explanatory power for SAT-10 performance in science and social studies. Sentence verification technique has also proven influential in predicting variance in models of reading comprehension, along with measures such as maze, oral reading fluency, and written retell (Marcotte & Hintze, 2009). With a host of measures demonstrating the ability to explain variance in achievement, there is a need to explore the efficacy of various combinations of measures in terms of their collective technical adequacy, instructional effectiveness, and logistical feasibility (Deno & Fuchs, 1987). Such inquiry would build on as well as possibly inform the research being conducted in middle grades reading progress monitoring assessment (e.g., Barth et al., 2012; Tolar, Barth, Fletcher, Francis, & Vaughn, 2014) and likely impact both general and special education. Special educators can contribute to inquiry in classrooms across the continuum of alternative placements, particularly inclusive general education classroom settings.

## References

Alexander, F. (n.d.). *Understanding vocabulary*. Retrieved from http://www.scholastic.com/teachers/article/understanding-vocabulary [Accessed 17 March 2013].

Barth, A. E., Stuebing, K. K., Fletcher, J. M., Cirino, P. T., Francis, D. J., & Vaughn, S. (2012). Reliability and validity of the median score when assessing the oral reading fluency of middle grade readers. *Reading Psychology*, *33*, 133-161.

Bloom, B. S. (1980). The new direction in educational research: Alterable variables. *Phi Delta Kappan*, *61*, 382-385.

Carney, R. N. (n.d.). Review of the Stanford Achievement Test Tenth Edition. *Mental Measurements Yearbook*. Retrieved from http://buros.org/mental-measurements-yearbook

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, *52*, 219-232.

Deno, S. L., & Fuchs, L. S. (1987). Developing curriculum-based measurement systems for data-based special education problem solving. *Focus on Exceptional Children*, *19* (8), 1-16.

Deno, S. L., Reschly, A. L., Lembke, E. S., Magnusson, D., Callender, S. A., & Windram, H. et al. (2009). Developing a school-wide progress-monitoring system. *Psychology in the Schools*, *46*(1), 46-55. doi: 10.1002/pits.20353.

Espin, C. A., Busch, T. W., Lembke, E. S., Hampton, D. D., Seo, K., & Zukowski, B. A. (2013). Curriculum-based measurement in science learning: Vocabulary-matching as an indicator of performance and progress. *Assessment for Effective Intervention*, *38*, 203-213. doi: 10.1177/1534508413489724.

Espin, C. A., Busch, T. W., Shin, J., & Kruschwitz, R. (2001). Curriculum-based measurement in the content areas: Validity of vocabulary matching as an indicator of performance in social studies. *Learning Disabilities Research & Practice*, *16*, 142-151. Retrieved from http://dx.doi.org/10.1111/0938-8982.00015

Espin, C. A., & Deno, S. L. (1994-1995). Curriculum-based measures for secondary students: Utility and task specificity of text-based reading and vocabulary measures for predicting performance on content area tasks. *Diagnostique*, *20*, 121-142.

Espin, C. A., & Foegen, A. (1996). Validity of general outcome measures for predicting secondary students' performance on content-area tasks. *Exceptional Children*, *62*, 497-514.

Espin, C. A., Shin, J., & Busch, T. W. (2005). Curriculum-based measurement in the content areas: Vocabulary matching as an indicator of progress in social studies learning. *Journal of Learning Disabilities*, *38*, 353-363. Retrieved from http://dx.doi.org/10.1177/00222194050380041301

Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review*, *33*, 188-192.

Fuchs, L. S., & Deno, S. L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional Children*, *57*, 488-500.

Jenkins, J. R., & Fuchs, L. S. (2012). Curriculum-based measurement: The paradigm, history, and legacy. In C. A. Espin, K. L. McMaster, S. Rose, & M. M. Wayman (Eds.). *A measure of success: The influence of curriculum-based measurement on education* (pp. 7-23). Minneapolis, MN: University of Minnesota Press.

Louisiana Department of Education. (LDE; n.d., a). *Integrated Louisiana Educational Assessment Program (iLEAP).* Retrieved from http://www.louisianaschools.net/lde/uploads/9725.pdf

Louisiana Department of Education. (LDE; n.d., b). *iLEAP 2010* (Technical Summary). Retrieved from http://www.louisianaschools.net/lde/uploads/18005.pdf

Marcotte, A. M., & Hintze, J. M. (2009). Incremental and predictive validity of formative assessment methods of reading comprehension. *Journal of School Psychology*, *47*, 315-335. doi: 10.1015/j.jsp.2009.04.003.

Moodle (n.d.). Retrieved from http://docs.moodle.org/23/en/About_Moodle.

Mooney, P., McCarter, K. S., Russo, R. J., & Blackwood, D. L. (2013). Examining an Online Content General Outcome Measure Technical Features of the Static Score. *Assessment for Effective Intervention*, *38* (4), 249-260.

Mooney, P., McCarter, K. S., Russo, R. J., & Blackwood, D. L. (2014). The structure of an online assessment of science and social studies content: Testing optional formats of a general outcome measure. *Social Welfare Interdisciplinary Approach*, *4* (1), 81-93.

Mooney, P., McCarter, K. S., Schraven, J., & Callicoatte, S. (2013). Additional performance and progress validity findings targeting the content-focused vocabulary matching. *Exceptional Children*, *80* (1), 85-100.

Morse, D. T. (n.d.). *Review of the Stanford Achievement Test* (10th ed.). *Mental Measurements Yearbook*. Retrieved from http://buros.org/mental-measurements-yearbook

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects*. Washington, DC: Authors.

Pearson Education (n.d.). *Stanford Achievement Test Series* (abbreviated form, 10th ed.). Retrieved from http://www.pearsonassessments.com/learningassessments/products/100000563/stanford-achievement-test-series-tenth-edition-abbreviated-battery.html

Royer, J. M. (2004). Uses for the sentence verification technique for measuring language comprehension. *Progress in Education*. Retrieved from http://www.readingsuccesslab.com/publications/Svt%20Review%20PDF%20version.pdf

Royer, J. M., Hastings, C. N., & Hook, C. (1979). A sentence verification technique for measuring reading comprehension. *Journal of Reading Behavior*, *11*, 355-363.

Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research. *Psychology in the Schools*, *42*, 795-819. doi: 10.1002/pits.20113.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson Education, Inc.

Tolar, T. D., Barth, A. E., Fletcher, J. M., Francis, D. J., & Vaughn, S. (2014). Predicting reading outcomes with progress monitoring slopes among middle grade students. *Learning and Individual Differences*, *30*, 46-57. doi: 10.1016/j. lindif.2013.11.001.

Vannest, K. J., Parker, R., & Dyer, N. (2011). Progress monitoring in Grade 5 science for low achievers. *The Journal of Special Education*, *44*, 221-233.

Wallace, T., Espin, C. A., McMaster, K., Deno, S. L., & Foegen, A. (2007). CBM progress monitoring within a standards-based system: Introduction to the series. *The Journal of Special Education*, *41*, 66-67.

Warne, R. T. (2011). Beyond multiple regression using commonality analysis to better understand R 2 results. *Gifted Child Quarterly*, *55*(4), 313-318.

Zientek, L. R., & Thompson, B. (2006). Commonality analysis: Partitioning variance to facilitate better understanding of data. *Journal of Early Intervention*, *28* (4), 299-307.

## VALIDITY OF TWO GENERAL OUTCOME MEASURES OF SCIENCE AND SOCIAL STUDIES ACHIEVEMENT

Paul Mooney, Renée E. Lastrapes,
Amanda M. Marcotte, Amy Matthews, B. S.

### Summary

Structured formative assessment in reading and mathematics has long been part of the fabric of special education. However, inquiry addressing the efficacy of structured formative assessment techniques to document student achievement and growth in social studies and science content is still in its infancy. Originally, reading measures were evaluated to determine their utility in measuring progress in the content areas. Over the past two decades the literature has expanded to include content focused instruments including vocabulary matching and content maze.

The present research addressed three more content-oriented instruments that have the advantage of being administered and scored online: Critical content monitoring, sentence verification technique, and written retell. The instruments were being evaluated for their technical adequacy and logistical

feasibility. Two research questions were evaluated: (a) what were the correlations with nationally-normed standardized achievement and statewide accountability tests in science and social studies content for critical content monitoring, sentence verification technique, and written retell? and (b) what was the incremental validity of adding measures of reading comprehension, using sentence verification technique and/or written retell, to content-focused achievement models that included critical content monitoring?

Participants were fifth-grade students (N = 51) enrolled in a public primary school in the southeastern U.S. As a whole, participants were 11.1 years old (SD = .5) at the time of testing, 68.6% female (n = 35), 66.7% Caucasian (n = 34), and 68.6% full-pay lunch status (n = 35). The three predictor variables were correlated with content test scores from the nationally representative standardized achievement test (i.e., Stanford Achievement Test-Tenth Edition abbreviated online form) and a statewide accountability test. Pearson correlations for critical content monitoring and the Stanford tests across science (r = .55) and social studies (r = .63) were moderately strong and similar in magnitude with other reported correlations for academic language measures in the literature. Correlations for critical content monitoring were descriptively larger than those between the standardized tests and sentence verification technique and written retell. Commonality analyses indicated that both critical content monitoring and sentence verification technique added unique variance to explanatory models.

A discussion of the results contributed to two implications. First, academic language, at the core of structured formative assessment instruments such as vocabulary matching and critical content monitoring, appears to be a viable avenue for continued inquiry. Second, given educational recommendations to rely on data from multiple sources in decision-making processes and present findings indicating that multiple instruments added unique variance to explanatory models, the use of multiple structured formative assessment measures in the development of content assessment frameworks appears warranted.

Study limitations included the order of presentation of the instruments, the small size and makeup of the sample, and the focus on one grade level.